Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
e-ISSN: 2395-3454

Vol. 18 | No. 3 | pp. 1077-1082 | 2019

Original Research Paper | Open Access

Detection of Sewage Discharge by Density Peak Search and Differential Expression Analysis

Weiguo Sun*†, Xudong Zhao and Hong Chen***

*College of Economics and Management, Northeast Forestry University, Harbin, 150040, China
**College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China
†Corresponding author: Weiguo Sun

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 22-12-2018
Accepted: 27-02-2019

Key Words:
Industrial sewage
Differential analysis
Fast density peaks
Clustering

ABSTRACT

Nowadays, water pollution is a part of the major environmental problems. Industrial sewage that does not meet the emission standards will pollute the surface water and groundwater when it is discharged into water bodies, causing serious adverse impacts on human beings and environment. In view of industrial sewage privately discharged without properly monitoring, we present a method for detection of sewage discharge using clustering and differential analysis on image sequences derived from satellite photos taken when focusing on a certain place. The proposed method helps to indicate key images containing sewage, make the sewage area and leave evidence for the retrospective incident. Clustering based on the search of fast density peaks is used for detecting images containing sewage. In addition, two sample's *t*-test and Fisher linear discriminant analysis are combined to extract the key pixels representing the area of sewage discharge. Experiments were made on 200 images corresponding to a certain area at different times of the day and 25 key frames with areas labelled to be sewage discharge were extracted, which indicated the effectiveness of this method.

INTRODUCTION

Unqualified industrial wastewater discharged into the environment may cause problems such as waterborne infectious diseases, toxic heavy metals in plants around water bodies, acute and chronic biological poisoning, etc. (Kocanova et al. 2017, Leiviska et al. 2017). It is particularly important to detect the illegal discharge of industrial sewage efficiently and in real time. Prevailing sewage detection methods include the method of field epidemiological investigation, questionnaire survey and discrimination on the cause of pollution by combining the detection results of source water, product water and terminal tap water before and after pollution, etc. (Prajapati et al. 2017). The above methods can accurately locate polluters and polluting areas, which will take a lot of manpower and material resources and cannot detect sewage in real time. Based on enterprise's sewage discharge system, monitoring in real time and on line can be made using industrial sewage monitoring systems developed by telephone network (Fu et al. 2017). It has the characteristics of safety and high cost performance, but it is impossible to detect the private and secret sewage discharges from factories. Therefore, the use of visual-based timing detection is to be considered.

In fact, the detection of key images of suspected sewage discharge from the obtained image set of the same scene belongs to the problem of image classification. Prevailing researches on image classification mainly use different indexing technologies of colour (Hu et al. 2017, Isaza et al. 2017), texture-based technologies (Liu et al. 2017, Haralick 1973), shape-based technologies (Masoumi et al. 2017, Liu 2017) and image classification technologies based on spatial relationship (Guo et al. 2017, Qin et al. 2017). All the above methods need to know the classification labels in advance. The image set used for sewage discharge detection keeps no classification labels, which makes the traditional classification methods lapse.

Using clustering technologies, we can get the frame containing suspected sewage discharge and label key areas of suspected sewage discharge on it. Thus, we propose a sewage discharge detection method based on clustering and differential expression. Clustering based on fast density peaks (Rodriguez et al. 2014) is introduced, and we used it on the satellite images acquired at the same area but at different times in order to classify the images into two groups, i.e., images with and without sewage discharge. Moreover, we pointed out key pixels representing sewage discharge on these images using combined two-sample *t*-test and Fisher linear discriminant criterion, which helps to leave evidence for retroactivity.

MATERIALS AND METHODS

Data source: In order to realize our sewage discharge detec-

tion, it is the most important issue on how to achieve clear satellite images. A website software served as a universal electronic map named as "Shui Jing Zhu (Commentary on the Waterways Classic)", which includes historical images, records clear images at the same location of different time. Using it, we downloaded 200 images at a section of river course in upper reaches of "Majia Gou" River in Harbin. Each image is processed with its resolution 400×300, which helps to validate the effectiveness of our method.

Research framework: First of all, preprocessing is made. Each image is pretreated by graying, enhancement, registration and vectorization. Secondly, we make clustering by fast search of density peaks on vectors representing images. The clustering result helps to decide whether private and secret sewage discharges from factories exist or not. Thirdly, a combined detection on key pixels representing sewage discharge is made by both two-sample *t*-test and Fisher linear discriminant criterion. Besides, morphological analysis was made, which helps to label the area representing sewage discharge. Corresponding research framework is shown in Fig. 1.

Preprocessing: First of all, colour images are made to gray ones to reduce the scale of data processing. Secondly, brightness enhancement is made for each gray image, which can treat overbright or too dark visible images and excessive concentration of gray scale on them, due to limited lighting. Here, we use the method proposed by Lavania & Shivali (2012) to implement the suppression of different illumination in the scene. Thirdly, registration is made on enhanced images. Harris corner is used. An enhanced image is selected as the standard image. Filtering operation using the difference operator is used for noise removal, and corners are obtained. Each image is adjusted according to the corner points obtained on it for registration. Fourthly, vectorization on each adjusted image is made. Each image transformed into a one-dimensional vector in a row-by-column fashion, and the position correspondence between two-dimensional image and one-dimensional vector is recorded. That is, the point (i, j) in an image is projected to the location $(i\text{-}1)\,\mathrm{col}\times j$ in a vector, where i and j denote the label of row and column respectively, and col represents the number of columns.

Image clustering: After vectorization, images constitute a two-dimensional expression profile, where the row and the column represent the location and the label of image. Clustering by fast search and find of density peaks (Rodriguez et al. 2014) is utilized across columns to accomplish clustering on images. According to the obtained decision graph, it can be recognized whether there is only one cluster or not. Once two cluster centres exist, one cluster may represent sewage discharge. Considering the probability of sewage images appearing in the whole images, we choose the cluster with smaller densities to represent the cluster of sewage discharge. Because of clustering in high-dimensional space, it needs a long learning time. After the learning period, images newly acquired and preprocessed can be quickly classified into either cluster by comparing its distances with the two cluster centres. Whether the sewage discharge exists or not can be promptly judged.

Labelling of pixels expressed differently: Once a definite clustering result is obtained, it can be determined whether there is a suspected illegal sewage discharge. Next, we concentrate on finding the key position denoting the difference between two kinds of images on two-dimensional expression profile derived from vectorized images, restoring them to the corresponding positions in two-dimensional image space and labelling the pixels representing key locations of suspected sewage discharge.

Permutations on two-sample *t*-test: In order to extract key pixels immediately, we ignore spatial correlations, and consider univariate enumeration for feature selection only. On each position component i of the two-dimensional expression profile, two-sample t-test is made. That is,

$$t(\upsilon(i)) = \frac{m_2(i) - m_1(i)}{\sqrt{\frac{s_1^2(i)}{n_1} + \frac{s_2^2(i)}{n_2}}}, \quad \ldots(1)$$

Where, *n1 and* n_2 represent the number of samples derived from two groups; $m_1(i)$ and $m_2(i)$ denote the luminance mean of the two groups; $S_1^2(i)$ and $S_2^2(i)$ correspond to variance; $v(i)$ denotes the degree of freedom. If there is no significant difference in the expression level between the two types of samples for feature i, then mixing up the classification labels will not significantly affect the calculation results of t-statistics. That is, multi-round scrambling of labels affect the calculation results of t-statistics only with random noise, which obeys normal distribution. Thus, t-statistics calculated without scrambling labels ought to be close to the mean of the distribution, which makes its p value approximate to 0.5. Otherwise, if p value is small enough, it may show that the calculation result of t-statistics after disturbing the label of samples is caused by the distribution of samples. In other words, there is a significant difference in the expression level between the two types of samples. The idea of permutation is equivalent to expanding the sample size, which makes the identified key positions with differentially expressed values more reliable and significant. Using permutation, the p-value calculation formula is as follows:

$$p_t(i) = \sum_{b=1}^{B} \frac{\#\{|t_0(i)|^3 |t(i)|\}}{B}, \quad ...(2)$$

Where, B denotes times for disturbing the label of samples; t_0 is the t-statistics after each round of disturbance. t represents the real t-statistics without any disturbance.

Permutations on linear discriminant analysis: In addition to distribution-based feature selection method for identifying key areas of suspected sewage discharge, we also consider using classification-based feature selection method, i.e., further verify the reliability of selected features with classification results. Fisher linear discriminant analysis (i.e., LDA) is one of the commonly used dimension reduction techniques for pattern classification (Duda et al. 2001). The equation of Fisher LDA classifier is as follows,

$$w^t x + w_0 = 0, \quad ...(3)$$

Where, w represents the normal vector; x is the sample vector; w_0 denotes the arithmetic mean of two mean values for two classes. That is,

$$w_0 = w(m_1 + m_2)/2, \quad ...(4)$$

Where, m_1 and m_2 represent the two mean values for two classes. When neglecting the quantity equilibrium of the two groups of samples, the classification error rate can be calculated as follows,

$$\begin{cases} Er_1 = \#\{w^t x + w_0 >= 0\}/n_1 \\ Er_2 = \#\{w^t x + w_0 < 0\}/n_2 \end{cases}, \quad c = (Er_1 + Er_2)/2 \quad ...(5)$$

Where, n_1 and n_2 represent the number of samples from the two groups. If the problem degenerates to the enumeration of univariables, the following judgment can be made without projection,

$$\begin{cases} w = 1, & \text{if } m_1 > m_2 \\ w = -1, & \text{else} \end{cases} \quad ...(6)$$

Then, we can get w_0 using equation (4), and can calculate the classification error rate using equation (5). Furthermore, we consider calculating the classification error rate by leave-one-out method, and compute the corresponding p value according to the thought of permutation, as is shown in equation (2). That is,

$$p_c(i) = \sum_{b=1}^{B} \frac{\#\{c_0(i)^3 c(i)\}}{B} \quad ...(7)$$

Joint enumeration for feature selection: Here, we consider both distribution-based and classification-based feature selection method, and obtain a pair of p values. That is (p_t, p_c) which corresponds to equation (2) and equation (7),

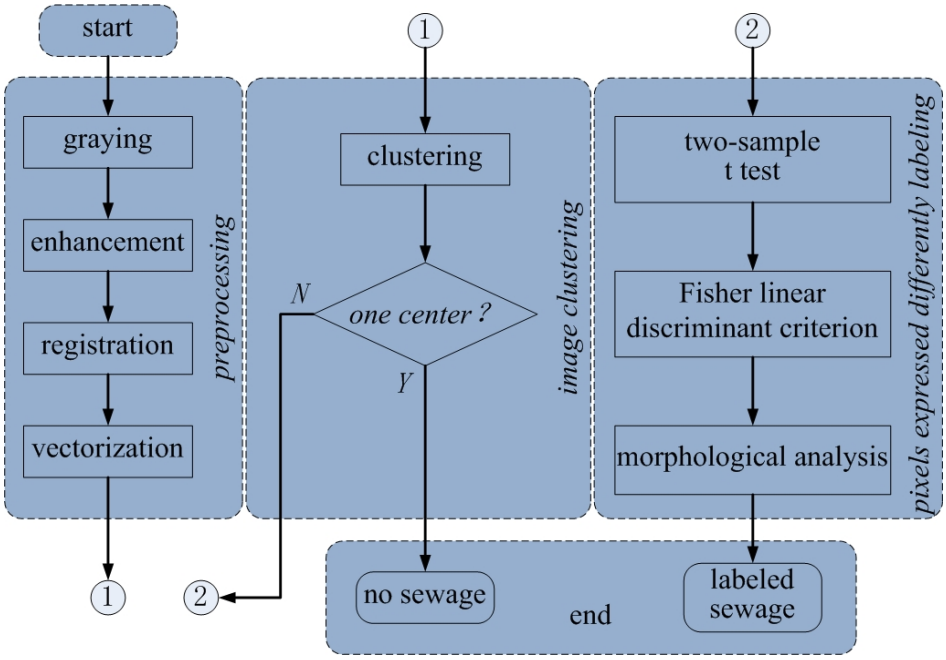

Fig.1: Framework of sewage discharge detection.

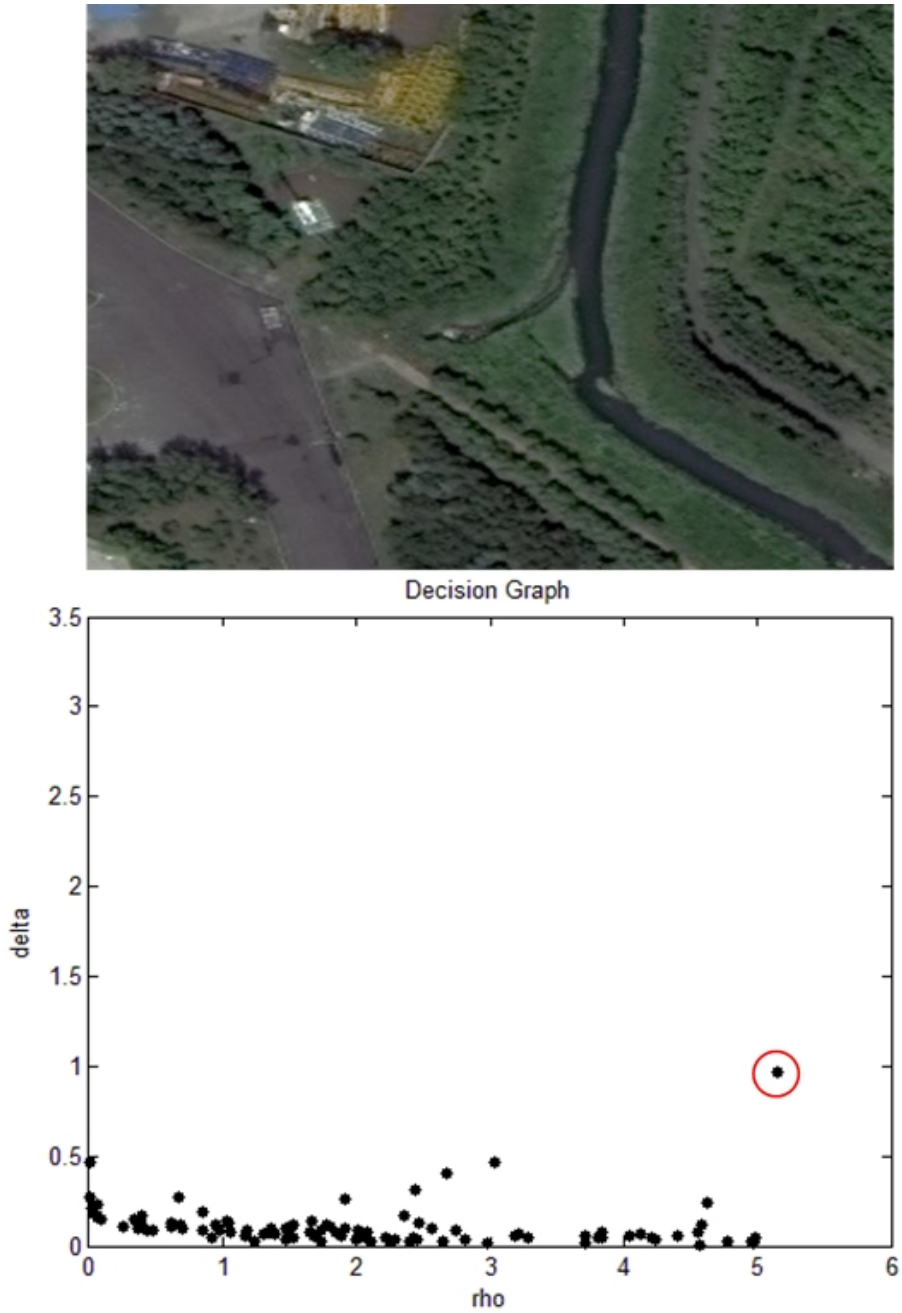

Fig. 2: Results of normal 100 images.

respectively. In fact, such combination using these two approaches needs to be made, in order to screen for relatively reliable feature. This kind of joint enumeration can be used to extract key pixels representing sewage discharge on images. We propose a simplified manual debugging scheme. That is, manual selection of thresholds for two types of p values is made for obtaining key pixels.

Image morphological analysis: Using joint enumeration for feature selection, we can select key pixels of suspected sewage discharge in the scene. In order to connect them into regions, we make a morphological analysis on obtained key pixels using erosion and dilation technology. Key pixels combine a complete region, gaps are filled, fine particles are eliminated and edges are smoothed. Finally, images la-

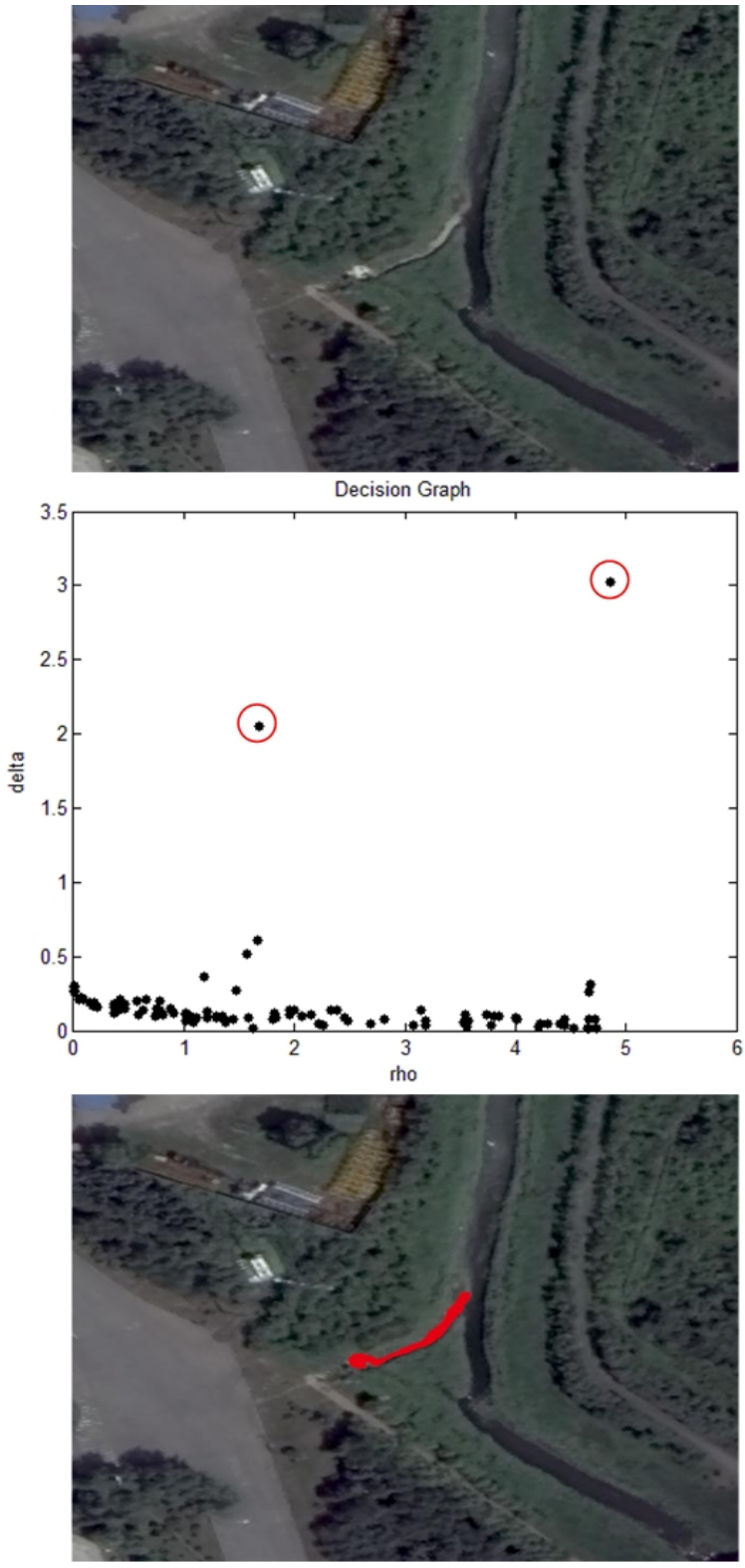

Fig. 3: Results of 100 images with suspected sewage discharge.

belled with sewage discharge area are obtained.

RESULTS AND ANALYSIS

To verify the effectiveness of the above methods, we used Awebsite software served as a universal electronic map named as "Shui Jing Zhu (Commentary on the Waterways Classic)", and downloaded 200 images at a section of river course in upper reaches of "Majia Gou" River in Harbin. Matlab R2012b was used to fulfil the framework of sewage discharge detection.

Fig. 2 shows the results of 100 images in a normal period. The upper subfigure is the scene of original image corresponding to the unique cluster center in this period. The lower subfigure is the decision graph on these 100 images using fast search on density peaks. It can be seen that there is a highest density point with farthest distance among each pair of two points (circled out in red), which indicates that the 100 images belong to the same kind of cluster. Thus, it can be concluded that there is no suspected sewage discharge during this period.

Fig. 3 indicates the results of 100 images during an abnormal period. The upper subfigure is the scene of original image corresponding to the cluster center with a lower density, which can be seen as a suspected sewage discharge. The middle subfigure is the decision graph on these 100 images using fast search on density peaks. It can be seen that there are two cluster centres (circled out in red). Based on clustering, a scene image representing the cluster center after differentially expressed analysis, is labelled on pixels for suspected sewage discharge detection, as shown in the lowest subfigure.

The above experimental results directly verify the effectiveness of the proposed framework as shown in Fig. 1. Indeed, it will take a long period of learning. Once the learning step is over, we can directly make a judgment on new images whether there is sewage discharge or not by comparing the distances between the vectorized images and the cluster centres.

CONCLUSION

A sewage discharge detection method based on clustering and differential expression is proposed in this paper. On the basis of traditional preprocessing techniques, we make clustering on images of a given scene during a period by fast search on density peaks, in order to judge whether there is suspected sewage discharge or not. Besides, we design a feature extraction method based on combined double-sample *t*-test and Fisher linear discriminant criterion for automatic labelling of sewage discharge area on images. The proposed framework helps to indicate the sewage images, mark the sewage area of the pictures and leave evidence for the retrospective incident.

REFERENCES

Duda, R.O., Hart, P.E. and Stork, D.G. 2001. Pattern Classification (2nd Edition). Wiley.

Fu, L., Huda, Q. and Yang, Z. 2017. Autonomous mobile platform for monitoring air emissions from industrial and municipal waste water ponds. Journal of the Air and Waste Management Association, 67(11): 1205-1212.

Guo, S., Huang, W. and Qiao, Y. 2017. Improving scale invariant feature transform with local color contrastive descriptor for image classification. Journal of Electronic Imaging, 26(1): 013-015.

Haralick, R.M. 1973. Texture features for image classification. IEEE Transactions on Systems Man and Cybernetics, 3(6): 610-621.

Hu, D.K., Zhang, L. and Zhao, W.D. 2017. Object classification via PCA-net and color constancy model. Applied Mechanics and Materials, 635-637: 997-1000.

Isaza, C., Anaya, K. and Paz, J.Z.D. 2017. Image analysis and data mining techniques for classification of morphological and color features for seeds of the wild castor oil plant. Multimedia Tools and Applications, 77(2): 2593-2610.

Kocanova, V., Cuhorka, J. and Dusek, L. and Mikulasek, P. 2017. Application of nanofiltration for removal of zinc from industrial wastewater. Desalination and Water Treatment, 75: 342-347.

Lavania, K.K. and Shivali, K.R. 2012. Image enhancement using filtering techniques. International Journal on Computer Science and Engineering, 4(1): 14-20.

Leiviska T., Khalid M.K. and Sarpola A. and Tanskanen, J. 2017. Removal of vanadium from industrial wastewater using iron sorbents in batch and continuous flow pilot systems. Journal of Environmental Management, 190: 231-242.

Liu, P.Z., Guo, J.M. and Chamnongthai, K. 2017. Fusion of color histogram and LBP-based features for texture image retrieval and classification. Information Sciences, 390: 95-111.

Masoumi, M. and Hamza, A.B. 2017. Spectral shape classification: A deep learning approach. Journal of Visual Communication and Image Representation, 43: 198-211.

Prajapati, D.R., Kaur N. and Prajapati D.R. 2017. Process capability in terms of TSS in waste water treatment technology. International Journal of Environment and Waste Management, 19(3): 203-215.

Qin, Y. and Tian, C. 2018. Weighted feature space representation with kernel for image classification. Arabian Journal for Science and Engineering, 43: 7113-7125.

Rodriguez, A. and Laio, A. 2014. Clustering by fast search and find of density peaks. Science, 344(6191): 1492-1496.