检测分析

DOI: 10.19902/j.cnki.zgyz.1003-7969.210740

基于红外光谱的食用植物油种类鉴别

孙一健,王继芬,张 震

(中国人民公安大学 侦查学院,北京 100038)

摘要:为建立基于红外光谱的食用植物油种类鉴别方法,收集了常见的5种食用植物油样本296份,采集红外光谱,分别通过Savitzky-Golay平滑、希尔伯特变换、IIR低通滤波器、IIR高通滤波器、连续小波变换、一阶导数、二阶导数进行预处理,并利用径向基函数(RBF)神经网络和随机森林(RF)模型对光谱进行识别。结果表明:RBF神经网络模型的效果优于RF模型,将红外光谱数据经希尔伯特变换处理后,RBF神经网络模型的识别率达到100%。采用该方法对食用植物油进行种类鉴别快速无损、准确率高、效果好。

关键词:食用植物油;鉴别;红外光谱;光谱预处理;径向基函数神经网络;随机森林

中图分类号:O657.33;TS225.1 **文献标识码**:A **文章编号**:1003-7969(2023)01-0120-05

Identification of edible vegetable oils based on infrared spectrum

SUN Yijian, WANG Jifen, ZHANG Zhen

(School of Investigation, People's Public Security University of China, Beijing 100038, China)

Abstract: In order to establish the identification method of edible vegetable oils based on infrared spectrum, 296 samples of 5 kinds of common edible vegetable oils were collected, their infrared spectrum were collected and pretreated by Savitzky-Golay smoothing, Hilbert transform, IIR low-pass filter, IIR high pass filter, continuous wavelet transform, first derivative and second derivative respectively, and spectrums were identified by Radial Basis Function (RBF) neural network and Random Forest (RF) models. The results showed that the effect of RBF neural network was better than the RF model. After pretreating the infrared spectrum data by the Hilbert transform, the recognition rate of the RBF neural network model reached 100%. This method has the advantages of rapid non-destructive, high accuracy and good effect in the identification of edible vegetable oils.

Key words: edible vegetable oil; identification; infrared spectrum; spectral pretreatment; Radial Basis Function neural network; Random Forest

近年来,随着生活水平的不断提高,人们对食品安全的重视程度也越来越高,其中,植物油作为膳食结构中不可缺少的重要组成部分,是食品安全的重要方面。有些不法商家为牟取暴利,对食用植物油进行掺假。2020年,公安部统一部署全国公安机关开展“昆仑行动”,严厉打击食品领域的犯罪[1],对于在食用植物油制假现场提取到的油痕,检验人员可以通过分析比对得到其种类等信息,为公安机关提供案件的调查方向,缩小侦查范围。

目前,对于植物油种类鉴别的方法有气相色谱-质谱法[2-3]、气相色谱-离子迁移谱法[4]、红外光谱法[5]等。红外光谱法具有分析快速、成本低、操作简单、无需样品预处理等优点。He等[6]将傅里叶变换红外光谱(FT-IR)与化学计量学相结合用于山茶油掺假的鉴定,其采用偏最小二乘判别分析的方法,分别基于不皂化物与植物油构建模型,成功鉴别了与山茶油脂肪酸组成相近,以及与山茶油脂肪酸组成不同的掺假山茶油。赵静等[7]以7个品种的77份合格植物油、28份不合格植物油以及118份地沟油为研究对象,使用二极管阵列近红外光谱仪采

收稿日期:2021-12-01;**修回日期**:2022-09-25

作者简介:孙一健(1997),男,硕士研究生,研究方向为理化检验(E-mail)154834607@qq.com。

通信作者:王继芬,教授(E-mail)wangjifen58@126.com。

集数据,采用多元方差分析以及贝叶斯判别分析对所采集的样品数据进行统计学分析,结果表明,贝叶斯判别函数模型对原始数据的分类准确率达 96.0%,交叉验证的准确率达 95.5%。

目前,关于红外光谱的研究大多集中在使用不同模型对食用植物油的种类进行鉴别,而对不同预处理方法的研究较少。在实际应用中,红外光谱数据信息存在噪声以及背景的干扰,使原始光谱的特征峰出现重叠、信噪比降低、基线漂移等情况。光谱预处理方法是指利用平滑处理、小波变换、滤波器、包络、抽取等方法减少由于仪器自身原因所导致的基线漂移等情况,消除红外谱图噪声和背景的干扰,从而提高模型对红外光谱的识别准确率,其中,常见的平滑算法包括 Savitzky - Golay、相邻平均法等,常见的小波变换算法包括连续小波、分解和重建小波以及多尺度离散小波等,常见的滤波器算法有 FFT 滤波器以及 IIR 滤波器等[8]。因此,在建模之前进行光谱预处理十分必要。

径向基函数(Radial Basis Function, RBF)神经网络是一种非线性 3 层静态的前馈式神经网络,包括输入层、隐藏层和输出层,从隐藏层到输出层的传递函数通常选取高斯函数[9-12]。随机森林(Random Forest, RF)模型是由若干个分类回归树进行预测的集成学习方法[13-14]。本文收集了 5 种常见的食用植物油,对其进行红外光谱采集,采用 7 种预处理方法对原始光谱进行处理,应用 RBF 神经网络以及 RF 模型方法建立预测模型对预处理后的红外光谱图进行识别,以识别率来比较不同预处理方法和不同模型对 5 种食用植物油分类的效果,为食用植物油的种类鉴别提供参考。

1　材料与方法

1.1　实验材料

296 份食用植物油样本,包括芝麻油 100 份、花生油 79 份、玉米油 37 份、亚麻籽油 40 份、橄榄油 40 份,基本信息见表 1。

表 1　296 份植物油样本的基本信息

植物油	数量(份)
芝麻油	
保定曲阳小磨香油	20
芝锦小磨香油	20
永溢粮油小磨香油	20
六必居小磨香油	20
金起小磨香油	20
花生油	
鲁花 5S 压榨一级花生油	20
第一坊冷榨花生油	20
山东烟台胡姬花古法花生油	20
胡姬花古法花生油	19
玉米油	
鲁花压榨特香玉米胚芽油	17
金龙鱼金滴玉米油	20
亚麻籽油	
罗尔仕压榨亚麻籽油	20
红井源压榨一级纯香亚麻籽油	20
橄榄油	
Olivoila 橄榄油	20
贝蒂斯特级初榨橄榄油	20

Nicolet is10 型傅里叶变换红外光谱仪,美国 Thermo Fisher Scientific 公司。

1.2　实验方法

1.2.1　光谱采集

在每个食用植物油样本上使用标签注明食用植物油的种类以及品牌,并进行编号。取 2 mL 食用植物油样放入石英样品杯中,然后放置于傅里叶变换红外光谱仪样品池中,盖上样品池的盖子,采集红外光谱。每个样品测量 3 次取平均值。

红外光谱仪参数:扫描次数 64 次,光谱分辨率 2 cm^{-1},测量范围 4 000 ~ 400 cm^{-1},动态调整 130 000 次/s,信噪比 50 000∶1 。

1.2.2　光谱预处理及因子分析降维

对采集到的红外光谱采用基线校正、平滑处理、滤波器 3 种方法进行预处理。其中:基线校正包括一阶导数微分(first derivative, FD)、二阶导数微分(second derivative, SD)和连续小波变换(Continue Wavelet Transform, CWT),导数处理选择基于 Norris 方法的导数处理,CWT 采用基于 Haar 类型的多尺度离散小波变换;平滑处理为 Savitzky - Golay(S - G)平滑;滤波器包括希尔伯特变换、无限冲激响应(Infinite Impulse Response, IIR)低通滤波器、IIR 高通滤波器,IIR 滤波器基于 Butterworth 方法进行处理。

预处理之后,将 296 份食用植物油样本的红外光谱信息转化为数据,采用 Z - score 的方法进行标准化处理,采用基于主成分分析为提取方法的因子分析方法对标准化后的数据进行降维处理,将所得到的特征变量用作后续建模分析。

1.2.3 食用植物油种类鉴别

应用 RBF 神经网络及 RF 模型方法建立预测模型对预处理后的红外光谱图进行识别，以识别率比较不同预处理方法和不同模型对 5 种食用植物油分类的效果。RF 模型算法中，设置最大树深度为 20，最大节点数为 10 000，其余数值均为 SPSS Modeler 系统的默认值。

2 结果与讨论

2.1 光谱预处理及因子分析降维结果

2.1.1 光谱预处理谱图的比较

296 份食用植物油的原始和预处理红外光谱图如图 1 所示。

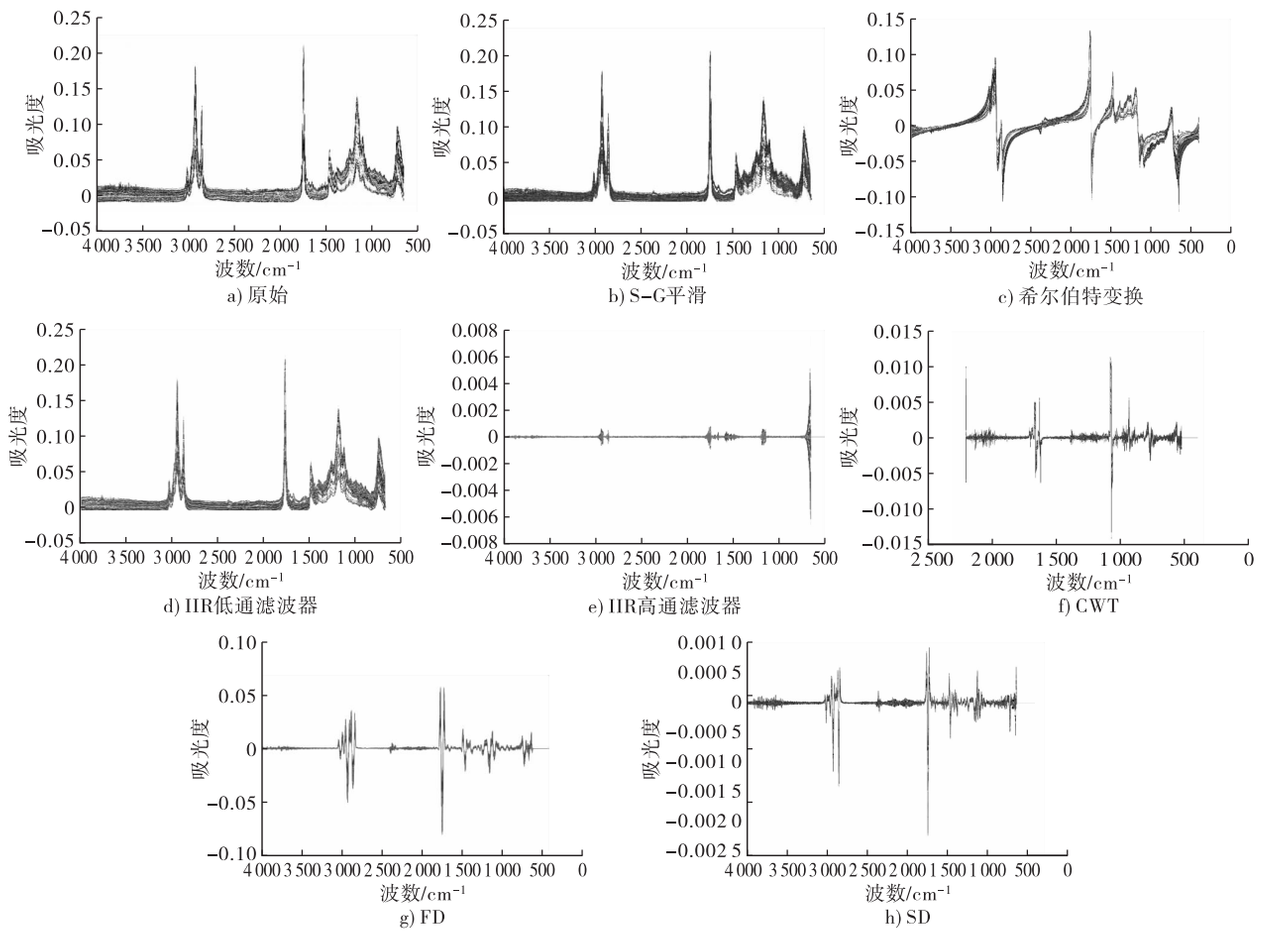


图 1 296 份食用植物油的原始和预处理红外光谱图

由图 1a 可知，2 900 cm^{-1}左右的尖强峰为 C—H 伸缩振动峰，1 750 cm^{-1}左右的尖强峰为 C═O 伸缩振动峰，1 200 cm^{-1}左右的中强峰为食用植物油中甘油三酯的 C—O 伸缩振动峰，1 450 cm^{-1}左右的弱尖峰为亚甲基的弯曲振动峰。不同种类的食用植物油具有相同或相似的吸收峰，但是出现了较为严重的重叠现象，同时受仪器条件以及采集环境的影响，出现了一定的基线漂移以及较为严重的背景干扰。

由图 1b～图 1h 可以看出，经过预处理之后，谱图的背景噪声有所降低，基线漂移现象也有所改善，各峰的区分度明显提高，但是各峰之间仍然存在相互交织的现象，通过肉眼很难进行准确区分，需要引入机器学习的方法实现对食用植物油红外光谱图的识别。

2.1.2 因子分析降维结果

不同预处理方法得到的红外光谱经过降维后，所提取的特征向量数各不相同，不同预处理方法的主成分数及累积方差贡献率见表 2。由表 2 可知，各种预处理方法的累积方差贡献率都能够达到 98% 以上，因此选择合适的主成分数可以实现特征数据对样本原始信息的保留。

表 2 不同预处理方法的主成分数及累积方差贡献率

预处理方法	主成分数	累积方差贡献率/%
无	15	99.501
S－G 平滑	35	99.046
FD	183	99.215
SD	206	99.021
CWT	254	98.307
IIR 高通滤波器	289	98.036
IIR 低通滤波器	45	99.014
希尔伯特变换	30	99.442

2.2 RBF 神经网络建模分析

使用 RBF 神经网络对经过因子分析降维后的特

征向量进行建模分析，在 SPSS Statistics 26 软件中设置随机取样来确定训练集和验证集的比例，5 种食用植物油在不同预处理方法下的识别率如表 3 所示。

表 3　5 种食用植物油在不同预处理方法下的识别率　%

预处理方法	芝麻油	花生油	玉米油	亚麻籽油	橄榄油
无	100	90	50	70	80
S－G 平滑	100	100	10	90	70
FD	100	100	38	47	80
SD	62	50	37	8	20
IIR 低通滤波器	75	100	100	100	90
IIR 高通滤波器	95	85	50	20	76
希尔伯特变换	100	100	100	100	100
CWT	87	100	100	100	88

由表 3 可知，5 种食用植物油在不同预处理方法下识别率具有较大差异。芝麻油和花生油在 SD 预处理方法下识别率较低，分别为 62% 和 50%，而在其他几种预处理方法下具有较好的识别率。玉米油和亚麻籽油在不同预处理方法下的识别率差异较大，其中在 IIR 低通滤波器、希尔伯特变换、CWT 预处理方法下识别率均达到 100%，而在其余几种预处理下，识别率较低。除 SD 预处理外，橄榄油在其他预处理方法下的识别率较为平均。从不同预处理方法来看，SD 预处理方法对 5 种样本的总体分类识别率不理想，可能是二阶导数在微分过程中丢失了部分原始数据，从而导致识别率大大下降，而经过希尔伯特变换预处理后，5 种样本的识别率均达到 100%。

希尔伯特变换下 5 种食用植物油样本的空间分布如图 2 所示，其中 RBF－1、RBF－2、RBF－3 表示在三维空间中食用植物油分类的特征轴。

由图 2 可以看出，5 种食用植物油样本被完全分离，表明该预处理方法下 RBF 神经网络的识别率达到最优。其中，花生油和亚麻籽油内部的分类很密集，表明这 2 种油的各品牌之间差异不大，而玉米油内部分类比较分散，表明不同品牌的玉米油之间差异较大。橄榄油和芝麻油之间分离较远，表明这两种油之间存在着较为明显的差异，能够被所建立的模型精确识别。

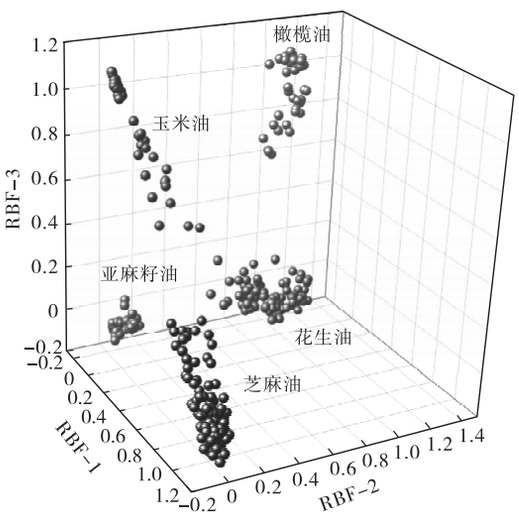
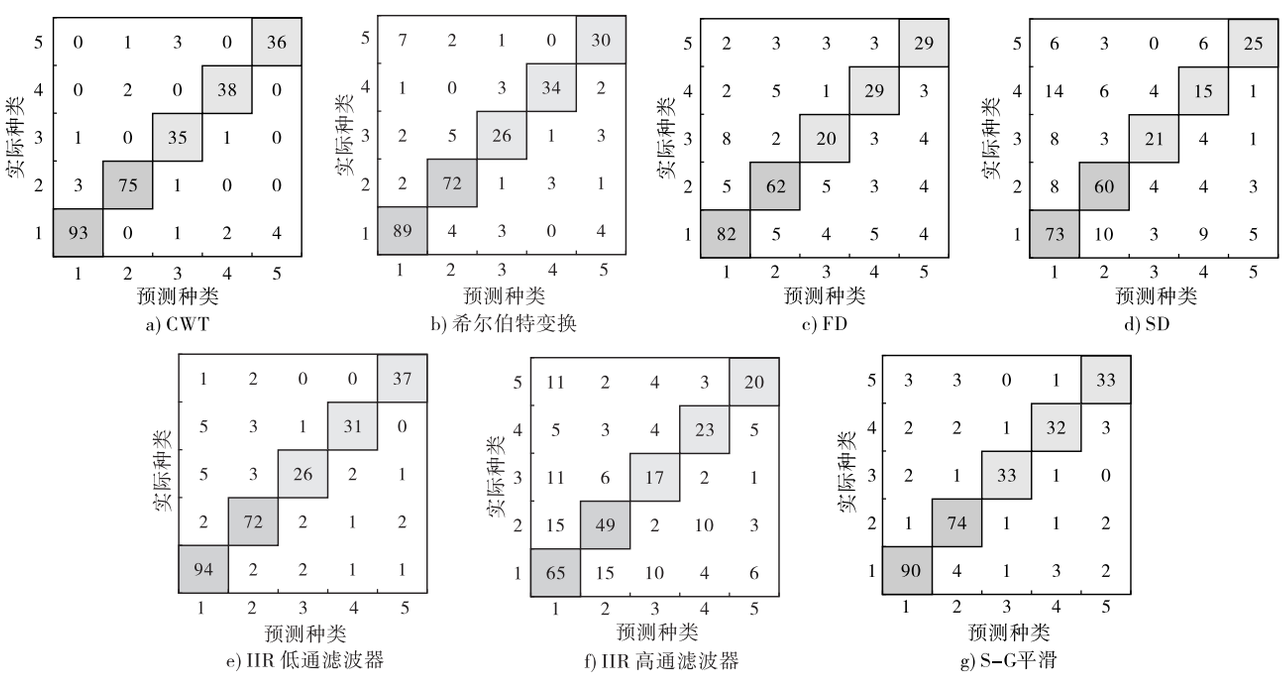


图 2　希尔伯特变化下 5 种食用植物油样本的空间分布

2.3　RF 建模分析

采用 RF 对经过因子分析降维后的数据进行建模分析，不同预处理方法下 RF 模型的混淆矩阵如图 3 所示。



注：1. 芝麻油；2. 花生油；3. 玉米油；4. 亚麻籽油；5. 橄榄油

图 3　不同预处理方法下 RF 模型的混淆矩阵

由图3可知,经过RF模型进行分类后,经CWT预处理的识别率最高,达到了94%,经S-G平滑预处理后的识别率达到89%,经FD预处理后的识别率为75%,经SD预处理后的识别率为66%,经希尔伯特变换预处理后的识别率为85%,经IIR低通滤波器预处理后的识别率为88%,经IIR高通滤波器预处理后的识别率59%。经IIR低通滤波器预处理后的识别率显著高于经IIR高通滤波器预处理后的,可能是样品中光谱信号主要是低频分量,IIR低通滤波器能够抑制光谱信号的高频分量而使光谱信号的低频分量通过,而IIR高通滤波器的处理相反,因此经过IIR低通滤波器预处理后的光谱数据总体上好于经过IIR高通滤波器预处理后的。经CWT预处理后的识别率最高,说明经过CWT预处理后,光谱数据中包含了绝大部分与食用植物油种类鉴别相关的信息,且与原始光谱数据相比,过滤了噪声等无用信息,同时RF算法利用了不同食用植物油种类之间小波变换中蕴含的变化,因而能够很好地对食用植物油的种类进行区分。

RF模型中经CWT预处理后的识别率(94%)低于在RBF神经网络模型下经希尔伯特变换预处理后的识别率(100%),分析原因可能是RBF神经网络模型的鲁棒性优于RF模型,其抗外界的干扰能力强,因此识别率高。

3 结 论

本研究利用傅里叶变换红外光谱技术,采集了5种食用植物油的光谱数据,通过不同的红外光谱预处理方法,结合RBF神经网络和RF建模,开展了食用植物油种类的鉴别。结果表明,RBF神经网络模型比RF模型更加适用于食用植物油的分类,在RBF神经网络模型中,对光谱进行希尔伯特变换预处理能够达到最高的识别率,识别率为100%。本研究为食用植物油种类鉴别提供了一种快速无损的新方法,该方法操作简单,准确率高,且无需昂贵的设备,十分利于在公安基层进行推广,为公安机关检验和分析食用植物油的种类提供了一定的参考。

参考文献:

[1] 孙一健, 王继芬. 太赫兹时域光谱技术在食品、药品和环境领域中的应用研究进展[J]. 激光与光电子学进展, 2022, 59(16): 22-31.

[2] 王同珍, 余林, 邱思聪, 等. 气相色谱-质谱技术结合化学计量学对6种植物油进行判别分析[J]. 分析测试学报, 2015, 34(1): 50-55.

[3] 鲍晓瑾, 倪炜华, 沈锡贤. GC-MS法识别二元混合植物油掺混量的方法研究[J]. 中国油脂, 2016, 41(12): 81-84.

[4] 陈通,陆道礼,陈斌. GC-IMS技术结合化学计量学方法在食用植物油分类中的应用[J]. 分析测试学报, 2017, 36(10): 1235-1239.

[5] 沈乐丞, 曾秀英, 温志刚, 等. 基于近红外光谱技术的赣南茶油掺假快速鉴别[J]. 中国油脂, 2022, 47(6): 62-67.

[6] HE W X, LEI T X. Identification of camellia oil using FT-IR spectroscopy and chemometrics based on both isolated unsaponifiables and vegetable oils[J/OL]. Spectrochim Acta A, 2019, 228: 117839[2021-12-01]. https://doi.org/10.1016/j.saa.2019.117839.

[7] 赵静, 梁瑞, 刘新保, 等. 近红外全波段扫描技术建立数学模型鉴别地沟油方法研究[J]. 中国油脂, 2021, 46(9): 71-76.

[8] 第五鹏瑶,卞希慧,王姿方, 等. 光谱预处理方法选择研究[J]. 光谱学与光谱分析, 2019, 39(9): 2800-2806.

[9] GOMES C R, MEDEIROS J A C C. Neural network of Gaussian radial basis functions applied to the problem of identification of nuclear accidents in a PWR nuclear power plant[J]. Ann Nucl Energy, 2015, 77: 285-293.

[10] HE Q, SHAHABI H, SHIRZADI A, et al. Landslide spatial modelling using novel bivariate statistical based Nave Bayes, RBF classifier, and RBF network machine learning algorithms[J]. Sci Total Environ, 2019, 663: 1-15.

[11] ZHANG P, ZHOU X, PELLOCCIONE P, et al. RBF-MLMR: a multi-label metamorphic relation prediction approach using RBF neural network[J]. IEEE Access, 2017, 5: 21791-21805.

[12] RIVERA A J, GARDA-DOMINGO B, JESUS M, et al. Characterization of concentrating photovoltaic modules by cooperative competitive radial basis function networks[J]. Expert Syst Appl, 2013, 40(5): 1599-1608.

[13] BREIMAN L. Random forests[J]. Mach Learn, 2001, 45: 5-32.

[14] BELGIU M, DRAGUT L. Random forest in remote sensing: a review of applications and future directions[J]. ISPRS J Photogramm, 2016, 114: 24-31.