加强科技创新支撑新时代防震减灾事业现代化建设
全国地震重点监视防御区公共服务 试点 工作通讯

2020 年第 17 期（总第 18 期） 2020 年 11 月 26 日

混淆矩阵评估的若干指标与地震预测预报评估讨论

张盛峰　张永仙　吴忠良

地震数值预测研究和传统方法评估试点项目
地震监测站网评估试点项目
人工智能地震监测分析系统完善与应用
地震危险区精细调查和地震现场综合科学考察试点项目
预报员访学试点项目
地震信息专题图试点

地震重点监视防御区公共服务试点

中国地震局地震预测研究所

加强科技创新支撑新时代防震减灾事业现代化建设
全国地震重点监视防御区公共服务 试点 工作通讯

2020 年第 17 期（总第 18 期） 2020 年 11 月 26 日

混淆矩阵评估的若干指标与地震预测预报评估讨论

张盛峰 张永仙 吴忠良

摘要

在目前的机器学习领域，存在较多的预测模型或算法，而与这些模型相对应的评估方法占有重要位置。只有明确不同评估方法的基本原理，才能更好地选择与模型对应的评估方法，以对模型在不断的选择和训练过程中进行优化。对于地震预测预报工作，如我国坚持多年的年度会商工作或大形势判定工作，主要目的是利用多种手段综合判定未来 1 年或 2~3 年时间内具有高概率发生强震的危险区或重点监视区，后续针对这些地区进行震情跟踪工作。在与此类真正的“向前”预测相对应的评估中，每一危险区或者与其相关的判定方法都可以被认为是判定地震是否发生的模型或分类器，混淆矩阵（Confusion Matrix）评估方法则主要针对此种类型模型的性能进行评估。混淆矩阵是机器学习中常见的用以评估预测模型效果的方法，为加深在评估检验工作中对这一概念的理解并以防误用，本文分析了该方法的基本原理及相关的多个评价指标，并以地震预测预报应用为例阐述了需要注意的问题。

1. 混淆矩阵

Confusion Matrix, 又被称为混淆矩阵或误差矩阵，是机器学习领域用于评价一个二分类预测模型性能好坏的常用评估方法 (周志华 , 2016)，已在地震学中得到了较多应用 (Beyreuther and Wassermann, 2008; Buscema and Ruggieri, 2011; Perrone et al., 2018; Galkina and Grafeeva, 2019; Gulia and Wiemer, 2019; Mousavi et al., 2019; Khan et al., 2020; Mignan and Broccardo, 2020; Rouet‐Leduc et al., 2020; Tehseen et al., 2020)，如 Gulia and Wiemer (2019) 利用混淆矩阵评估了“红绿灯”形式的分类器在地震发生过程中对地震序列属于前震还是余震进行实时识别的能力；Mignan and Broccardo (2020) 利用混淆矩阵分析了机器学习中的人工神经网络（artificial neural network，ANN）模型与其他简易模型在预测增益上的不同；Rouet‐Leduc et al. (2020) 根据混淆矩阵和受试者工作特征曲线（Receiver Operating Characteristic curve，简称 ROC 曲线）的原理，分析了利用机器学习识别出来的背景噪声和震颤与实际情况的差别；Mousavi et al. (2019) 利用基于机器回归学习方法的地震监测器对地震事件进行检测，并利用混淆矩阵中的查准率、查全率和 F-Score 指标对模型性能进行了评估。为了便于对这一系列指标的理解，我们用年度危险区的预测检验来进行说明。每年年度会商会给出预测年份可能发生强震的多个年度地震危险区 (石耀霖和刘杰 , 2000; 孙其政和吴书贵 , 2007)，在预测时段内发生的强震与危险区的关系有四种：(1) 危险区内发生强震，可以表述为有震报准（TP）；(2) 强震发生在危险区外，可以表述为漏报（FN）；(3) 危

* 支撑新时代防震减灾事业现代化建设试点任务之一“地震数值预测研究和传统方法评估试点项目”成果。

险区内未发生强震，可以表述为虚报（FP）；（4）危险区外未发生强震，可以描述为无震报准（TN）（马宏生等，2004）。图 1 则用混淆矩阵（类似二位列联表 (Everitt, 1992)）来表述这四种情况 (Nwanganga and Chapple, 2020)。在这个矩阵中，四种情况分别为：真正（TP）、真负（TN）、假正（FP）和假负（FN）。

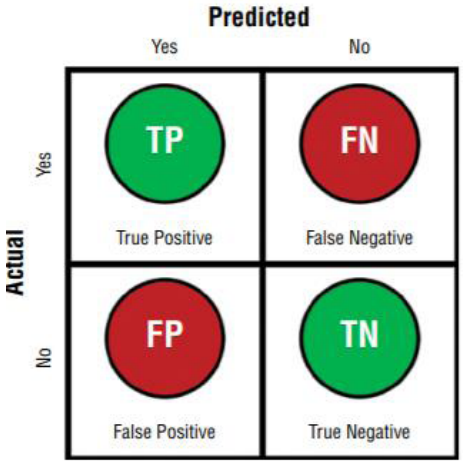

图 1 用以表示实际情况与预测情况的混淆矩阵示例 (Nwanganga and Chapple, 2020)

2. 准确率（accuracy）、错误率（error）

在划定的危险区对下一年地震发生情况的预测中，若想知道真正发生地震的危险区占总体多大的比例，此时需要用到第一个指标—Accuracy，称为准确率或正确率，是预测结果与实际情况相一致的概率，即 Accuracy=(TP+TN)/(TP+TN+FP+FN)。同样，错误率（Error）指二者不一致的概率，即 Error=(FP+FN)/(TP+TN+FP+FN)。准确率是较为直接的评价指标，但往往在实际工作中不够科学全面，常常准确率高并不能代表一个模型好，比如对于地震活动水平较低的地区，若某一模型只给出无震的预测意见，则准确率也很高（虽 TP 很低，但 TN 很高），因为此地区实际发生较大地震的概率本来就很低。准确率和错误率是分类任务中最为常用的两种性能度量，既可用于二分类任务，也可用于多分类任务，但无法满足所有任务需求，此时需要引入其他的性能度量。

3. 精确度或查准率（precision）、召回率或查全率（recall）或敏感度（sensitivity）

Precision，又称为精确度或查准率，为预测为正的样本中实际为正的情况所占比例，即 Precision=TP/(TP+FP)，指标意义为回答“预测有震的危险区中到底有多少发生了地震”的问题。反过来讲，若想知道实际发生地震的危险区中预测会发生地震的情况所占的比例，即回答“发生地震的危险区被找到了多少”，则需要用召回率或查全率（Recall）这个指标，即 Recall=TP/(TP+FN)，该指标也被称为敏感度（Sensitivity）。两种指标示意图如图 2 所示。这两种性能度量常用于信息检索、网络搜索等应用中，例如我们会经常关心“检索出的信息有多大比例是用户感兴趣的”(查准率) 和“用户感兴趣的信息中有多少被检索出来了”(查全率) 等问题。可见，查准率指标基于的样本总体是模型预测为正的情况（TP+FP），查全率指标基于的总体样本是实际发生为正的情况（TP+FN）。

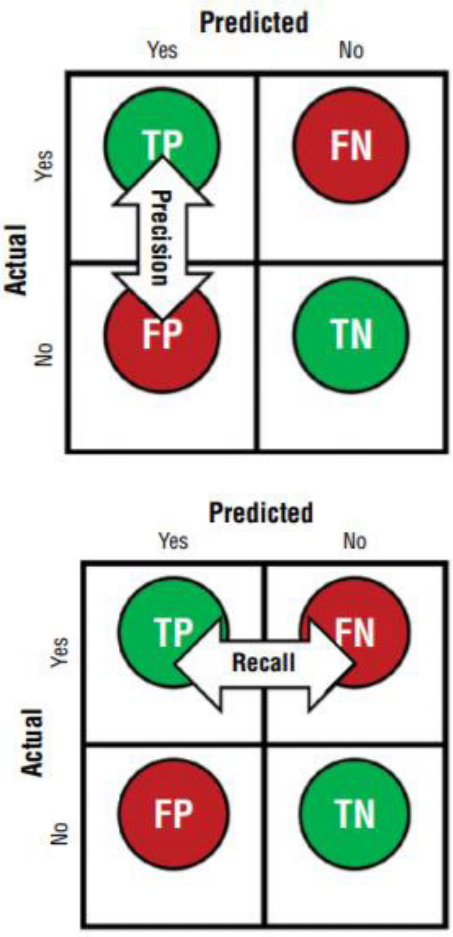

图 2 混淆矩阵中的查准率（上图）和查全率指标（下图）

4. R 值评分、敏感度、特异性

与查准率和查全率相对应，若将样本总体设定为实际发生情况分别为正 (TP+FN) 和否（FP+TN) 的情况，则会引入地震预测预报领域常用来评估预报效能的 R 值评分方法。R 值评分由许绍燮院士在 20 世纪 70 年代提出（许绍燮，1973），后续又进行了进一步修正（许绍燮，1989），分析原理与受试者工作特征分析方法 (ROC)(Swets, 1973; Holliday et al., 2005) 类似，且基本原理已被中国地震学家和国际学者广泛采用 (Keilis-Borok and Kossobokov, 1990; 石耀霖和刘杰，2000; 张国民等，2002; 马宏生等，2004; 朱令人和王琼，2004)。R 值评分定义为 R= 报准地震数 / 全部地震数 - 虚报面积 / 全部无震面积 = 报准率 - 虚报率或 R= 无震报准率 - 漏报率 =1- 漏报率 - 虚报率 = 报准率 + 无震报准率 -1，即：R 值 =TP/(TP+FN)-FP/(FP+TN)。若将地震学与机器学习领域进行对比，则报准率与敏感度对应，漏报率与 1- 特异性对应，即 Sensitivity=TP/(TP+FN)，Specificity=TN/(TN+FP)，如图 3 所示。

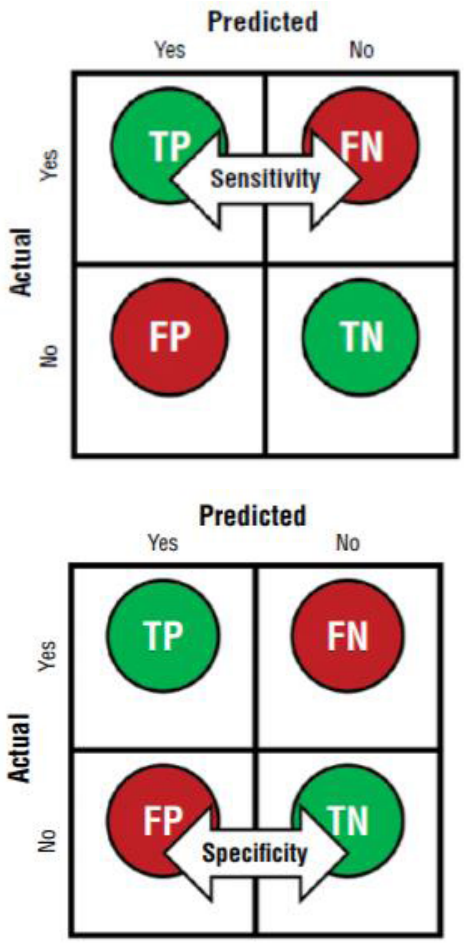

图 3 混淆矩阵中的敏感度（上图）和特异性指标（下图）

5. F_1 和 F_β 值

查准率和查全率是一对矛盾的量，一般而言，查全率高时，查准率偏低；而查准率偏高时，查全率往往不高。可以用以下两种场景形象的加以解释：

（1）“宁可错挑一千，也不放过一个”。要想提高查全率，只需要把可能要发生地震的情况全部预测为发震即可，更极端点儿，把大陆所有地方都标记为要发生地震，此时未来发震的地方全部包含在了我们的预测中，但同时，由于很多地方实际没有地震，而导致查准率偏低。

（2）“宁缺毋滥”。若想提高查准率，也就是我们预测有震的地方都会发生地震，这是我们所追求的效果，即通过各种地震前兆、手段去判定最有可能发生地震的地方。相比随机预测，这种做法确实让我们取得了一定成果，即查准率得到了提高，但是，由于很多地震前的指示指标并不明显，因此很容易从专家的高标准筛选中“漏掉”，导致查全率不会很高。

因此可以看出，仅考虑查准率或查全率都难以满足需求，我们需要综合考虑二者的情况，由此出现 F1 指标，即查全率和查准率的调和平均值，

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} = \frac{2 \times \mathrm{TP}}{\mathrm{Total} + \mathrm{TP} - \mathrm{TN}}$$

在某些情况下，需要对查全率和查准率附加不同的权重，则此时需要用到其一般形式 F_β，即查准率和查全率的加权调和平均值，

$$F_\beta = \frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall} = \frac{(1+\beta^2)\mathrm{TP}}{(1+\beta^2)\mathrm{TP} + \beta^2 \mathrm{FN} + \mathrm{FP}}$$

可以看出，当 β =1 时，F_β 退化为 F_1；当 β >1 时查全率有更大影响；当 β <1 时查准率有更大影响。通常在一些简单的预测任务中，才可能出现查准率和查全率都很高的情况。

6. P-R 曲线（Precision Recall Curve）

为了综合评价整体结果，很多情况下可根据模型的预测结果对样本数据进行排序，按此顺序逐个把样本作为正例进行预测，则每次可计算出当前的查准率和查全率。以查准率为 Y 轴、查全率为 X 轴形成的曲线称为 PRC，即“ P-R 曲线”，显示该曲线的图称为“ P-R 图”。它能直观地显示出预测模型在样本总体上的查全率和查准率，如图 4 所示，可以看出查准率随查全率的升高而降低。在比较两种预测模型的结果时，若一个模型的 P-R 曲线被另一种模型的曲线包围，则说明后者的性能优于前者，如图中的 A 模型优于 C；若两条 P-R 曲线发生交叉，如图中模型 A 和 B 的情况，此时两种模型在查准率和查全率两个指标上均较高，则较难说明哪一种模型效果更好，此时可引入另一指标——“平衡点”（Break-Even Point，简称 BEP）来衡量，即表示查准率和查全率相等的点。从图 4 中可以看出，模型 A 性能优于模型 B。

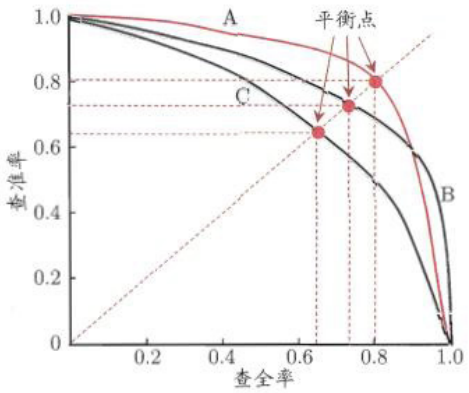

图 4 P-R 曲线与平衡点示例

（为显示方便和美观，此图显示出单调平滑曲线，但现实任务中的 P-R 曲线常是非单调、不平滑的，在局部存在很大波动的情况）

7、ROC 曲线

受试者工作特征 ROC（Receiver Operating Characteristic）曲线是地震预测模型效能的常用检验方法，它源于“二战”中用于敌机检测的雷达信号分析技术，后来被用于一些心理学、医学检测应用中，此后被引入机器学习模型或算法的评估中 (Spackman, 1989)。通过按照一定顺序设定阈值，预测模型可以给出不同阈值情况下的性能指标。根据地震预测领域中的惯例，

常以击中率（TPR=TP/(TP+FN)，同查全率或敏感度）为 X 轴、虚报率（FPR=FP/(TN+FP)，同 1-Specifity）为 Y 轴形成的曲线表示 ROC 曲线，其对角线对应于“随机预测”模型，如图 5 所示。

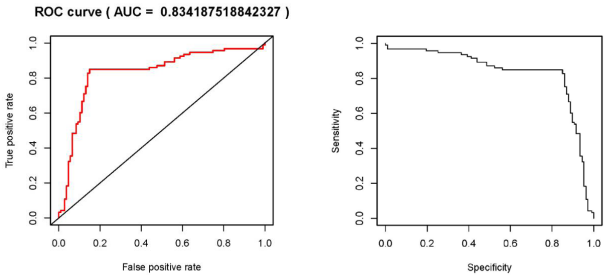

图 5 ROC 曲线示例（左图）及对应的敏感度 - 特异性曲线（右图）

8. AUC

与 P-R 曲线类似，若一个模型的 ROC 曲线被另一模型的 ROC 曲线完全包围，则说明后者模型的性能优于前者，而若出现两条曲线交叉的情况，则难以判断两种模型孰优孰劣。此时，较为合理的判据是计算 ROC 曲线以下的面积，即 AUC（Area Under Curve）值。假设 ROC 曲线上面点的坐标依次为 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$，其中 $x_1=0$，$x_m=1$，则 AUC 的值可估算为

$$AUC=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})$$

AUC 值越高说明模型的性能越好，若 AUC 为 1，则模型的性能近乎完美，当然这样的模型几乎不存在；当 AUC 等于 0.5，则模型的性能与随机预测类似；若 AUC 小于 0.5，则说明模型性能不如随机预测效果好；若两种模型得到的 AUC 值相同，说明总体上两种模型的性能类似，如图 6 所示。

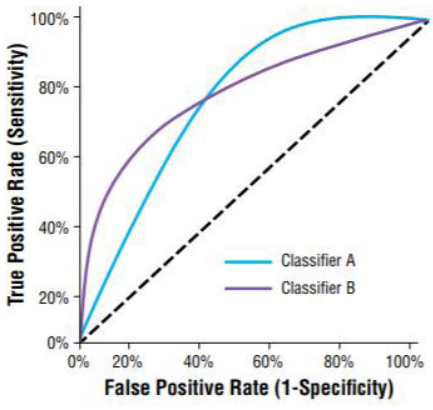

图 6 两种性能相同的模型（A 和 B）的 ROC 曲线

9. 在地震预测评估中的应用讨论

上述介绍了与混淆矩阵相关的若干性能指标，包含准确率和错误率、精确度（或查准率）和召回率（或查全率、敏感度）、R 值评分、特异性、F1 和 F β 值、P-R 曲线、ROC 曲线和 AUC 值。由于不同模型所面对的任务需求不同，需要根据不同的应用场景合理使用这些指标，才能得到科学有效的评估结果。同时，在地震预测预报领域，同样也存在不同的任务需求，比如常规业务工作中短临应急预测和对首都圈等特殊地区地震趋势的研判中，力求不放过每一个可能要发生的地震，因此对查全率要求较高，而针对中长期预测工作或强震活跃但致灾程度较低的地区来说，需要利用多种模型在一定时空尺度上给出较为准确的预测，因此较为倾向于“宁缺毋滥”的理念，对查准率要求较高。在实际情况中，除了一些预测模型可以给出“0”或“1”的预测结果，一些概率预测模型往往给出的是 0~1 之间的概率数值，利用概率密度函数和数据分布的理论同样也可将这些指标用于预测模型间的性能评估。

混淆矩阵给出的四象限情况包含预测与实际相对应的四种情况，而对于实际的地震业务工作，模型或方法给出的预测结果和混淆矩阵评估的需求并不完全吻合，如年度会商划定的危险区或针对不同构造区域进行的地震大形势判定，给出的结果更多的是针对 6.5 或 7 级地震预测为“yes”的情况，而较少主动的给出预测为“No”的情况，因此，在以往主要针对概率预测模型的评估检验中，预测为“No”的概率常以 1 减去预测为“yes”的概率表示。因此，从这个角度出发，与以往把焦点关注于计算某一地区在未来一段时间内发生强震的概率相比，若某一方法或模型能够主动计算得到在相同的时空范围内不发生地震的概率，这一工作不仅会满足后续利用混淆矩阵进行模型性能评估的四象限输入需求，同时也可为区域震情形势判定等工作提供重要的科技支撑。

参考文献

[1]Beyreuther M, Wassermann J. 2008. Continuous earthquake detection and classification using discrete Hidden Markov Models. Geophysical Journal International 175: 1055-1066.

[2]Buscema M, Ruggieri M. 2011. Advanced Networks, Algorithms and Modeling for Earthquake Prediction. River Publishers.

[3]Everitt B. 1992. The Analysis of Contingency Tables. New York: Chapman and Hall/CRC.

[4]Galkina A, Grafeeva N. 2019. Machine learning methods for earthquake prediction: A survey, in: Litvinov, Y.andTrifonov, P. (Eds.), Proceedings of the FourthConference on Software Engineering and Information Management (SEIM 2019)(pp. 25-32). (CEUR Workshop Proceedings; Vol. 2372). RWTH Aahen University.

[5]Gulia L, Wiemer S. 2019. Real-time discrimination of earthquake foreshocks and aftershocks. Nature 574: 193-199.

[6]Holliday J R, Nanjo K Z, Tiampo K F, et al. 2005. Earthquake forecasting and its verification. Nonlinear Processes in Geophysics 12: 965-977.

[7]Keilis-Borok V I, Kossobokov V G. 1990. Premonitory activation of earthquake flow: algorithm M8. Physics of the Earth and Planetary Interiors 61: 73-83.

[8]Khan I, Choi S, Kwon Y-W. 2020. Earthquake Detection in a Static and Dynamic Environment Using Supervised Machine Learning and a Novel Feature Extraction Method. Sensors (Basel) 20: 800.

[9]Mignan A, Broccardo M. 2020. Neural Network Applications in Earthquake Prediction (1994–2019): Meta-Analytic and Statistical Insights on Their Limitations. Seismological Research Letters 91: 2330-2342.

[10]Mousavi S M, Zhu W, Sheng Y, et al. 2019. CRED: A Deep Residual Network of Convolutional and Recurrent Units for Earthquake Signal Detection. Scientific reports 9: 10267.

[11]Nwanganga F, Chapple M. 2020. Practical Machine Learning in R. Wiley.

[12]Perrone L, De Santis A, Abbattista C, et al. 2018. Ionospheric anomalies detected by ionosonde and possibly related to crustal earthquakes in Greece. Annales Geophysicae 36: 361-371.

[13]Rouet-Leduc B, Hulbert C, McBrearty I W, et al. 2020. Probing Slow Earthquakes With Deep Learning. Geophysical Research Letters 47.

[14]Spackman K A. 1989. Signal detection theory: valuable tools for evaluating inductive learning, Proceedings of the sixth international workshop on Machine learning. Morgan Kaufmann Publishers Inc., Ithaca, New York, USA, pp. 160–163.

[15]Swets J A. 1973. The Relative Operating Characteristic in Psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. Science 182: 990-1000.

[16]Tehseen R, Farooq M S, Abid A. 2020. Earthquake Prediction Using Expert Systems: A Systematic Mapping Study. Sustainability 12.

[17] 马宏生 , 刘杰 , 吴昊 , 李杰飞 . 2004. 基于 R 值评分的年度地震预报能力评价 . 地震 2: 31-37.

[18] 石耀霖 , 刘杰 . 2000. 对我国 90 年代年度地震预报的评估 . 中国科学院研究生院学报 17: 63-69.

[19] 孙其政 , 吴书贵 . 2007. 中国地震监测预报 40 年 . 北京：地震出版社 .

[20] 许绍燮 . 1973. 震兆分析一例 . 地震技术资料汇编 . 北京 : 科学出版社 : 20-23.

[21] 许绍燮 . 1989. 地震预报能力评分 . 国家地震局科技监测司编 . 地震预报方法实用化文集地震学专辑 . 地震出版社 , 北京 , pp. 586-589.

[22] 张国民 , 刘杰 , 石耀霖 . 2002. 年度地震预报能力的科学评价 . 地震学报 15: 550-558.

[23] 周志华 . 2016. 机器学习 . 清华大学出版社 .

[24] 朱令人 , 王琼 . 2004. 新疆地震年度趋势预报效能的统计评价 . 内陆地震 4: 289-299.

加强科技创新支撑新时代防震减灾事业现代化建设
全国地震重点监视防御区公共服务 试点 工作通讯目录

关于观测仪器中的模拟滤波与数字滤波	2020 年第 1 期（总第 1 期）
推进新时代地震预测研究现代化框架方案（2020-2035 年）	2020 年第 2 期（总第 2 期）
2020 年 6 月 26 日新疆于田 M_S6.4 地震虚拟科学考察试点工作报告	2020 年第 3 期（总第 3 期）
研究所加强科技创新支撑新时代防震减灾事业现代化建设试点行动方案（2020 ～ 2022 年）	特刊第 1 期（总第 4 期）
地震预测基础研究成果支撑引领地震预测业务的若干基本问题	2020 年第 4 期（总第 5 期）
地震监测预报预警科技进展和发展趋势	2020 年第 5 期（总第 6 期）
地震危险区精细调查与地震现场综合科学考察规划（初稿）	2020 年第 6 期（总第 7 期）
北京地区活动断裂与地震图	2020 年第 7 期（总第 8 期）
科学规划地震预测的进步	2020 年第 8 期（总第 9 期）
中国地震科学实验场地震科学考察工作预案（初稿）	2020 年第 9 期（总第 10 期）
预测所地震重点监视防御区公共服务试点工作方案	2020 年第 10 期（总第 11 期）
地震大形势科学问题清单	2020 年第 11 期（总第 12 期）
人工智能实时地震监测分析系统的应用	2020 年第 12 期（总第 13 期）
亚太经合组织地震科学合作项目 ACES	2020 年第 13 期（总第 14 期）
地震电磁短临监测手段评估—GPS TEC	2020 年第 14 期（总第 15 期）
2021-2030 年中国大陆地震重点监视防御区确定工作报告	2020 年第 15 期（总第 16 期）
人工智能研究进展分析报告（2017-2020）	2020 年第 16 期（总第 17 期）
混淆矩阵评估的若干指标与地震预测预报评估讨论	2020 年第 17 期（总第 18 期）

编辑部：

中国地震局地震预测研究所科研管理部
E-mail:sycglb@ief.ac.cn