在日冕物质抛射中，哪些上游太阳风条件对预测 B_z 最重要

Pete Riley[1]，**M. A. Reiss**[2,3]**和 C. M¨ostl**[3]

[1]Predictive Science Inc.，美国加利福尼亚州圣迭戈。[2]社区协调建模中心，674 代码，美国国家航空航天局戈达德太空飞行中心，格林贝尔特，马里兰州 20771，美国[3]奥地利空间天气办公室，奥地利格拉茨中央气象与地球动力学研究所

arXiv:2303.17682v1 [physics.space-ph] 30 Mar 2023

要点：

- 由上游鞘区内特征驱动的机器学习算法可以准确预测抛射物中 B_z 的最小值
- 最重要且统计显著的特征是鞘层数密度和总场强。
- 这些特征反映了行星际日冕物质抛射上游的压缩情况，而这种情况与抛射物的整体磁场强度相关。

通讯作者：Pete Riley，pete@predsci.com

摘要

准确预测行星际磁场的 z 分量，尤其是在行星际日冕物质抛射（ICME）经过期间，是空间天气预报的一项重要目标。目前提出的技术寥寥无几，而且其适用范围和准确度仍然有限。近期，研究人员开发了一种稳健的机器学习（ML）技术，根据 42 个“特征”（即根据行星际日冕物质抛射上游及其鞘区内测得的量计算得出的变量），预测行星际日冕物质抛射内 B_z 的最小值。在本研究中，我们更详细地考察这些所谓的解释变量，重点关注：（1）统计显著的变量；以及（2）最重要的变量。我们发现，数密度和磁场强度解释了很大一部分变异性。这些特征反映了行星际日冕物质抛射对前方背景太阳风的压缩程度。直观而言，这合乎情理：日冕物质抛射喷发时可获得的能量会分配为磁能和动能。因此，能量更强的日冕物质抛射会以更强的磁通绳磁场（更大的 B_z）和更高的速度发射，从而造成更强的鞘层压缩（数密度和总场强增加）。

通俗语言摘要

随着社会对技术的依赖日益加深，准确预报地磁暴强度的需求也愈发重要。由快速日冕物质抛射（CME）引发的风暴构成了最大的威胁；有记录以来的所有重大地磁事件都与这类风暴有关，部分原因是它们内部的磁场最强。快速日冕物质抛射穿过太阳风时，会在其前方形成所谓的“鞘区”。本研究将鞘区的性质与日冕物质抛射磁场的强度联系起来，表明对压缩程度的估算能够最稳健地预报随后出现的磁场。这种关系为一种颇具前景的预报方法奠定了基础，可在日冕物质抛射内部峰值磁场抵达前一天以上发出预警。

1 引言

准确预报大尺度行星际磁场，尤其是其 z 分量， B_z 是任何空间天气预报系统的关键参数；然而，迄今为止，这一问题仍基本未能解决（一个例证是，目前没有任何预报中心提供针对 B_z 的预报）。多种过程都会使行星际磁场（IMF）的 z 分量不为零。然而，在没有瞬变效应的情况下，大尺度宁静螺旋状日球磁场的净 B_z 为零。波动和湍流可以叠加在这一大尺度图景之上（例如 Horbury & Balogh，2001年），但这些涨落本身并不会主动引发显著的空间天气影响。从地球效应的角度看，产生连贯磁通绳结构的大型太阳爆发在相对不受扰动的情况下传播至 1 天文单位，是地磁暴的主要来源，尤其是当磁通绳轴位于黄道面内或附近时。此外，快速日冕物质抛射会在其前方驱动快模激波，压缩行星际磁场并放大波动/湍流涨落。此外，大尺度磁场缠绕在抛射物周围，可产生较大且持续的 B_z 值（Gosling & McComas，1987年）。因此，从按优先级排序的实际运行角度来看，我们需要重点关注那些速度快、结构连贯且能形成强鞘区的日冕物质抛射（Lugaz等人，2016年）。

多年来，人们开发了多种颇有前景的方法来预测 B_z，从纯粹的统计技术（例如 Riley 等（2017）；J. Chen 等（1996））到自洽的、基于物理的全球磁流体动力学（MHD）模拟（例如 Riley 等（2007）；Shiota 和 Kataoka（2016）；T¨or¨ok 等（2018））。介于这两种极端之间的是一系列基于经验的方法，它们结合观测值与有限的经验或统计建模，以优化预测能力（例如 Savani 等（2015）；M B_z ostlet 等（2017））。

建模，以优化预测能力（例如 Savani 等（2015）；M¨ostlet 等（2017））。

机器学习（ML）技术已在许多科学领域得到广泛应用（例如，Witten 和 Frank（2005））。尤其是太阳与日球层研究，目前正受益于监督式技术和无监督式技术的应用，以便更好地理解，或至少描述和预测各种各样的现象（例如，Qahwaji 等（2008）；Bobra 和 Couvidat（2014）；Camporeale 等（2017、2018）；Heidrich-Meisner 和 Wimmer-Schweingruber（2018）；Bailey 等（2020）；Camporeale（2020年））。尤其是监督式机器学习算法，在解决回归问题方面可能是有效的方法。较为直接的多元回归方法仍然适用；然而，随机森林算法等更复杂的技术有望缓解过拟合、超参数调优以及处理缺失数据等问题（例如，Xia 等（2017））。不过有一点需要注意：尽管简单多元回归等较简单的算法容易理解、可自行编写，或至少可以仔细检查，但许多机器学习算法必须被视为“黑箱”。使用包含相关程序且已在许多科学领域经过广泛测试的标准软件包，可以让人确信算法实现是正确且稳健的。

在先前的一项研究中，我们应用机器学习方法，评估上游原位行星际日冕物质抛射（ICME）鞘区的测量值能否估算随后 ICME 中的（1）磁场 B_z 分量的最小值；或（2）总磁场（$|B|$）的最大值（Reiss 等，2021）。我们基于 Wind、STEREO-A 和 STEREO-B 航天器观测到的 348 个行星际日冕物质抛射，建立了预测模型。我们发现，目标变量与一组由 42 个输入（解释）变量或特征构成的集合之间存在较强关联（皮尔逊相关系数，PCC 分别为 $= 0.71$ 和 0.91）。这些解释变量由磁场三个分量、总磁场强度、等离子体密度、温度和太阳风整体速度的各种统计特征组成，具体包括：均值、标准差、最小值和最大值、最大值与最小值之比，以及均值与标准差之比（即变异系数）。因此，对于这七个变量中的每一个，都有六项统计测量值，共计 42 个特征或输入变量，理论上可以解释观测到的抛射物内 B_z 最小值或磁场最大值的变化。由于目标是建立预测模型，我们将 348 个事件分为训练数据集（4/5）和评估数据集（1/5），并应用了三种模型：线性回归器（LR）、随机森林回归器（RFR）和梯度提升回归器（GBR）。较高的 PCC 值表明，这可能是一种很有前景的预测工具，可在 ICME 抵达 1 天文单位前数小时估算其磁场强度。然而，原研究仍有几个问题未能解答。第一，42 个输入变量中哪些最重要？也就是说，哪些变量导致了 ICME 内 B_z 的大部分变化？第二，哪些变量具有统计显著性？第三，从科学角度看，至少部分输入变量与目标变量之间存在强关联的潜在机制是什么？

本研究旨在通过回答以下问题，在前期研究的基础上进一步推进。具体而言，我们希望找出哪些变量具有统计显著性，哪些变量对目标的变异性贡献最大，并解释其原因。我们认为，这也回应了日球物理学领域在应用机器学习技术时经常受到的质疑：我们能从中学到什么？尽管前期研究（Reiss 等，2021）直接回应了业务空间天气需求，因此并未将这些问题作为主要关注点，但本研究旨在运用机器学习方法，更好地理解为何会出现如此高的相关性。

在接下来的章节中，我们首先介绍数据集并概述其预处理方式，然后总结计划采用的机器学习技术。接下来，我们将详细描述一组

分析，并概述基于不同数据初始处理方式进行的若干次重新分析。最后，我们结合对行星际日冕物质抛射演化的直观认识来解读结果，讨论本研究的局限性，并提出未来改进本研究的建议。

2 方法

2.1 数据

Reiss 等（2021）从 HELIO4CAST 行星际日冕物质抛射目录（M¨ostl 等人，2020年）中识别出 364 个产生上游鞘区的行星际日冕物质抛射。要求这些事件在前方形成鞘层（或者更准确地说，表现出密度增强），这很可能意味着这些行星际日冕物质抛射的行进速度快于其所处的背景太阳风，不过压缩也可能由较慢抛射物的膨胀造成（Siscoe 和 Odstrcil, 2008；Salman 等人，2021）。所有事件均由地基观测设备或 STEREO-A/B 航天器在2007年1月至2021年3月期间观测到，因此都位于约 1 天文单位处。经过进一步筛选，最终保留了 348 个数据覆盖良好的事件。

此外，尽管原始目录中没有记录激波是否先于日冕物质抛射出现，我们将 HELIO4CAST 行星际日冕物质抛射目录与一份基于 Wind 数据集、记录由日冕物质抛射驱动的激波的独立目录（Nieves-Chinchilla 等人，2018年）进行了交叉比对，并创建了一个新事件数据集，其中以 0/1 标记是否存在激波。在全部 348 个事件中，有 247 个被发现驱动了行星际激波；另外 101 个事件则未发现明显相关的激波。

根据这些事件的时间序列，为以下各变量计算了六种统计量：均值、标准差（std）、最小值（min）、最大值（max）、最大值与最小值之比（minmax），以及变异系数（cv），即均值与标准差之比：总磁场（B_t）、磁场的三个分量（B_x、 B_y、 B_z）、整体等离子体速度（v_t）、数密度（n_p）和质子温度（T_p）。矢量量采用日心地球赤道坐标系（HEEQ）表示，该坐标系以太阳自转轴为基准：z 轴平行于太阳自转轴（正方向指向北方）；从地球观测时，x 轴指向太阳赤道与太阳中央子午线的交点；y 轴则构成右手正交坐标系。用于计算这些量的时间边界为鞘层区间的起点至终点。Reiss 等（2021）还研究了在每个时间窗口起点额外加入四小时数据的影响，因此解释变量中包含了少量实际抛射物数据。

该数据集（包含 348×42 (14616) 个数据点）汇总于表1中，表中第1列列出了各项特征。对于每项特征，表中给出了均值、标准差、最小值、第25和第75百分位数以及最大值。请注意，这些是特征的统计量，而特征本身也属于统计量。因此，“St. Dev.”（第4列）中“std.bx”的值，是对348个鞘区分别计算标准差后所得数值的标准差。最后一行汇总了“目标”或输出变量；在本例中，它是抛射物内 B_z 的最小值。缩写“std”指标准差，“cv”指变异系数，即均值与标准差的比值；“minmax”指最大值与最小值的比值。为方便起见， bt 表示总磁场分量 $|B|$，其他变量则很容易理解。

尽管我们稍后还会用到这些数值，但值得注意的是，在某些情况下，它们与背景太阳风的平均属性相差很大。例如，鞘区的数密度均值（np.均值）为 14.1 cm^{-3}，而

Table 1. List of parameters used in the multiple-regression analysis, together with their basic statistical properties. See the text for a detailed explanation of each parameter.

Feature	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
mean.bx.	348	−0.292	2.737	−8.388	−2.093	1.383	11.411
max.bx.	348	7.114	4.736	−4.941	3.838	9.357	32.120
std.bx.	348	3.239	1.914	0.000	1.946	4.206	11.714
min.bx.	348	−7.522	4.609	−27.209	−10.101	−4.387	3.577
cv.bx.	348	12.911	130.146	0.000	0.843	3.735	2,416.322
minmax.bx.	348	−1.745	15.220	−265.063	−1.506	−0.692	78.281
mean.by.	348	0.472	4.909	−21.306	−2.847	3.688	16.800
max.by.	348	10.184	6.908	−10.915	6.109	13.074	43.805
std.by.	348	4.591	3.270	0.000	2.536	5.908	25.036
min.by.	348	−9.401	7.095	−50.495	−12.904	−4.934	5.030
cv.by.	348	3.519	8.915	0.000	0.567	2.633	107.937
minmax.by.	348	−1.049	9.981	−138.319	−1.277	−0.535	120.411
mean.bz.	348	0.229	3.109	−9.976	−1.591	1.966	13.673
max.bz.	348	10.256	6.910	−5.495	5.821	13.033	60.193
std.bz.	348	4.558	2.968	0.000	2.667	5.790	20.432
min.bz.	348	−9.799	6.736	−40.842	−12.236	−5.743	7.602
cv.bz.	348	19.078	213.734	0.000	1.251	5.463	3,962.469
minmax.bz.	348	−0.957	2.311	−22.105	−1.209	−0.710	32.596
mean.bt.	348	9.399	4.546	2.026	6.417	11.183	28.042
max.bt.	348	14.362	7.992	2.681	8.777	18.131	61.447
std.bt.	348	2.253	1.741	0.000	1.122	2.901	14.293
min.bt.	348	3.076	2.274	0.228	1.535	4.115	16.107
cv.bt.	348	0.230	0.105	0.000	0.155	0.284	0.734
minmax.bt.	348	0.255	0.188	0.014	0.105	0.355	1.000
mean.vt.	348	447.167	106.012	283.544	369.984	494.505	823.933
max.vt.	348	498.517	131.658	300.348	401.584	567.684	1,001.504
std.vt.	348	19.926	15.332	0.000	8.860	27.290	119.767
min.vt.	348	400.165	88.781	239.981	337.344	439.992	723.750
cv.vt.	348	0.043	0.028	0.000	0.023	0.055	0.208
minmax.vt.	348	0.816	0.092	0.408	0.756	0.885	1.000
mean.np.	348	14.102	9.264	0.451	7.685	18.491	84.160
max.np.	348	28.930	20.587	1.164	14.502	37.137	147.326
std.np.	348	4.704	4.021	0.000	1.970	6.105	26.247
min.np.	348	5.812	4.904	0.112	2.728	7.271	49.334
cv.np.	348	0.326	0.173	0.000	0.201	0.422	1.135
minmax.np.	348	0.248	0.166	0.002	0.127	0.343	1.000
mean.tp.	348	153,878.500	150,688.400	13,725.420	52,988.160	207,266.300	964,935.600
max.tp.	348	560,333.500	751,663.600	19,754.490	117,018.800	647,930.000	5,249,402.000
std.tp.	348	74,633.040	96,655.200	0.000	15,796.020	91,670.020	659,472.700
min.tp.	348	45,772.240	44,346.380	4,565.096	20,026.440	54,835.390	360,332.500
cv.tp.	348	0.414	0.207	0.000	0.270	0.522	1.548
minmax.tp.	348	0.167	0.136	0.004	0.066	0.248	1.000
Target	348	−9.456	6.681	−60.782	−11.760	−5.506	0.000

投稿至 *Space Weather* 的稿件

超过 1 天文单位处背景太阳风平均数密度的两倍。此外，其最大值（84.2 cm^{-3}）高出一个数量级以上。

为评估数据范围变化对模型结果的影响，我们采用了几种方法对这些数据重新归一化。这有助于减轻部分特征取值范围过宽而对结果产生不成比例影响所导致的潜在偏差。首先，采用最小-最大归一化方法，将所有变量重新缩放，使最小值为零、最大值为一。虽然这种方法可能使所有数据处于相同范围内，但若存在异常值，反而可能导致某些数据集被压缩到不成比例的较小范围。因此，我们还采用了第二种方法，即Z分数重新归一化。这种转换是指对数据集中的每个值重新归一化，使所有值的均值为0、标准差为1。该方法既能处理数据范围过宽或过窄的问题，也能减轻异常值的影响。其唯一较小的影响体现在视觉解读方面：最小-最大归一化方法的优点是可以将所有数据呈现在0到1之间。除表1外，所有统计模型结果均以Z分数重新归一化为基准报告。

自原始研究完成以来（Reiss 等，2021），HE-LIO4CAST ICME 目录的更新版已在线发布（ICMECATv2.1）。该目录包含 521 个带有鞘区的事件，数据来自 Wind 1995—2021年以及 STEREO-A/B 2007—2021年。为了更直接地将本研究与原始研究联系起来，我们主要分析原始数据集中的事件。不过，我们也使用 v2.1 数据重复了分析，以验证新增事件并未改变主要推论和结论。

2.2 模型

本研究使用*R*中的多个软件包分析多元回归模型，主要采用“线性模型拟合”套件（lm）（R Core Team，2019）。不过，这些回归技术只能确定变量的统计显著性。为了深入估算各解释变量在描述变异性方面的相对重要性，我们采用了传统统计方法和较新的机器学习（ML）方法。

为确定能得到性能最佳模型（即降低预测误差的模型）的解释变量，我们使用了 MASS 软件包中的“stepAIC”函数（Venables和Ripley，2002年）。我们进行了前向选择和后向选择（即后向消除）（James等人，2013年）。在前向选择中，模型最初不包含任何预测变量，算法会逐步加入贡献最大的变量，直至改进不再具有统计显著性。

采用多种方法有助于评估变量排序所应具有的不确定性，因为不同技术依据不同的重要性指标。所用技术包括：随机森林（Liaw 等人，2002年）、XGBoost（T. Chen 和 Guestrin，2016年）、相对重要性（Gr¨omping 等人，2006年）、基于 earth 的多元自适应回归样条（MARS）（封装器，2019年）、逐步回归（Bendel 和 Afifi，1976年）以及 DALEX（Biecek，2018年）。选择这些方法，是因为它们代表了应用最广泛的方法。

需要再次强调的是，这些方法对“重要”的含义采用了不同的定义，因此，我们并不期望各项结果完全一致。尽管如此，结果一致之处最值得我们信赖；而结果不一致之处，则需要我们更加谨慎。

值得指出的是，我们选择采用广泛的机器学习技术，以突出这些方法的稳健性及其局限性。因此，将它们结合起来，

可以将其视为一个“模型集合”：方法间的一致性可用于增强我们的信心，而不一致性则会降低信心。另一方面，尝试利用不同方法的实现构建任何形式的集合预报框架，超出了本研究的范围。

我们还要指出，此处的目的是应用这些统计模型，了解这些解释变量中哪些对于解释抛射物内部 B_z 的变化最为重要。具体而言，这并非像 Reiss 等（2021）所做的预测性研究；我们的目标是理解机器学习技术为何表现如此出色，以及它们让我们了解到哪些因素驱动了这些关联。

3 结果

对 42 参数数据集进行多元线性回归分析后，发现五个解释变量最为显著（$p < 0.05$）：mean.bt.、std.bt.、cv.bt.、mean.np. 和 max.np.（表 2）。较低的 p 值表明，我们可以拒绝这样的零假设：解释变量与目标（ICME 内的 B_z）之间的关系在统计上不显著。总体而言，拟合优度指标 R^2 为 0.579，因此这些变量共同能够解释 ICME 内 B_z 变化的 58%。

我们特别关注这些变量，并重复了回归分析。由此得到表 3 所示的结果。据此，我们推断有五个变量在 $p < 0.01$ 水平上具有统计显著性（mean.bt.、std.bt.、cv.bt.、mean.np. 和 max.np.）。总体而言，我们推断，与 $|B|$ 和 n_p 相关的参数似乎具有统计显著性。直观来看，这说得通：$|B|$ 和 n_p 在压缩区内应当具有相关性。

我们可以直观地考察这些具有统计显著性的变量之间的关系。图 1 展示了这五个显著性最高的输入（解释）变量与目标变量的散点图矩阵。沿第一列向下或沿第一行横向查看，我们注意到，ICME（目标）中的 B_z 与五个输入变量均呈负相关。也就是说，较大的负 B_z 值与较大的正输入变量值相关。不过，我们也注意到散点分布较为分散，这一点也反映在置信区域中。此外，各输入变量之间均呈正相关，相关程度因参数而异，但总体而言，在 $|B|$ 或 n_p 分组内的相关性较高。比较这两组后，相关性较低。

前面的分析以启发式方法探讨了解释变量可能对描述 B_z^{ICME} 变化所作的贡献。接下来，我们采用多种算法运用更稳健的技术，评估统计显著性和重要性。在这些方法中，我们迭代地添加或删除解释变量，以确定能够产生最佳模型的变量子集，也就是预测误差最低的模型。

首先，我们使用 R 的 MASS 软件包，分别采用前向选择和后向消除：前向选择会迭代添加变量，直至改进不再具有统计显著性；后向消除则会迭代删除变量，直至所有变量均达到统计显著性（表 4）。根据这项分析，22 个变量可解释 B_z^{ICME} 中 57% 的变异性。这与我们通过逐一检索各变量的临时 方法计算出的 58% 大致相同。在此案例中，按重要性排序，统计显著性最高的参数依次为：标准差、np.均值、bt 变异系数、组间均值和最大 np。

使用随机森林方法评估变量的重要性后，我们发现最重要的八个变量（按重要性排序）是：np.均值、最小值（np）、

表2。表1所述42个解释变量的多元回归分析汇总。为简洁起见，部分被发现不具有统计显著性的条目已省略。括号中的项表示标准误差（即系数标准偏差的估计值），而残差标准误衡量回归模型对数据集的拟合程度，F统计量则检验线性回归模型中回归系数的显著性。

	Dependent variable:
	目标
max.bx.	0.047
	(0.114)
std.bx.	0.136
	(0.131)
最小值.bx.	0.056
	(0.111)
变异系数. bx.	−0.064
	(0.039)
按最小值和最大值的 bx	−0.045
	(0.038)
按均值	0.011
	(0.077)
按最大值	0.092
	(0.136)
按标准差	−0.200
	(0.141)
按最小值	0.118
	(0.151)
按变异系数	0.011
	(0.041)
按组最小最大值	0.036
	(0.038)
…	
均值	0.591***
	(0.196)
组间最大值。	−0.481
	(0.295)
标准差。	−0.775***
	(0.260)
组间最小值	−0.083
	(0.132)
bt 变异系数	0.648***
	(0.141)
…	
np.均值	−0.689***
	(0.194)
最大 np	0.440**
	(0.182)
std.np.	0.012
	(0.192)
最小值（np）	−0.083
	(0.131)
变异系数（np）	−0.183*
	(0.111)
…	
最小-最大值（tp）	0.148*
	(0.083)
常数	−0.000
	(0.037)
观测值	348
R^2	0.579
调整后 R语言^2	0.523
残差标准误	0.691（自由度 = 306）
F统计量	10.280*** （自由度 = 41；306）
Note:	*p<0.1； **p<0.05； ***p<0.01

按最小值、按标准差、组间最大值、mean.tp、max.bz. 和组间均值（表5）。除此之外，相对重要性有所下降。我们注意到，基于 bz 的变量可能会被视为

表 3. 用于预测行星际日冕物质抛射中 B_z 最小值的五个最显著解释变量汇总。

	因变量：
	目标
均值。	0.341^{***}
	(0.108)
标准差	-1.034^{***}
	(0.142)
bt 变异系数	0.522^{***}
	(0.094)
np.均值	-0.645^{***}
	(0.079)
最大 np	0.321^{***}
	(0.087)
常数	-0.000
	(0.039)
观测值	348
R^2	0.473
调整后 R语言^2	0.466
残差标准误	0.731（自由度 = 342）
F统计量	61.496^{***}（自由度 = 5；342）
注：	*p<0.1； **p<0.05； ***p<0.01

与 $|B|$ 在压缩区内有关，因此在某种程度上可以将它们视为混杂变量。将这份列表与之前识别出的显著变量进行比较，可以看出，任何一组显著且且 重要的解释变量至少都包括：均值、组间最大值和 mean.np。需要注意的是，反复运行随机森林方法会得到不同的估计值，因为这是一种采用随机变量的统计方法。因此，尤其是在依次考察这些变量时，对其相对排名的置信度会逐渐降低。例如，cv.tp. 和 minmax.tp. 的相对重要性经常互换；不过，它们各自与 mean.np 的比值始终显著小于 1。

使用 XGBoost 方法按重要性对解释变量进行排序，识别出的变量依次为：按最小值、组间最大值、mean.np、组间均值、按标准差、最小值（np）、按均值和 mean.tp（表 6）。除此之外，其余变量的影响力均不超过最重要变量的 10%。同样值得注意的是，与 $|B|$、 n_p 和 B_y 相关的特征占据了主导地位。

多元自适应回归样条（MARS）模型也可用于对解释变量进行排序。这是一种灵活的技术，此处采用该方法是为了说明我们的结果对所实施技术的敏感性——

表 4. 使用 MASS 软件包进行逐步回归。

	Dependent variable:
	目标
std.bx.	0.117^{*}
	(0.065)
cv.bx.	-0.068^{*}
	(0.038)
按最小值	0.247^{***}
	(0.076)
最小 Bz。	-0.195^{**}
	（0.078）
Bz 最大值与最小值。	0.058
	（0.038）
均值	0.591^{***}
	(0.144)
组间最大值	-0.468^{**}
	(0.198)
标准差	-0.777^{***}
	(0.170)
bt 变异系数	0.596^{***}
	(0.098)
vt.均值	0.452^{***}
	(0.172)
vt.标准差	-0.436^{*}
	(0.254)
vt.变异系数	0.472^{**}
	(0.221)
np.均值	-0.642^{***}
	(0.147)
最大 np	0.471^{***}
	(0.127)
最小值（np）	-0.171^{**}
	(0.081)
变异系数（np）	-0.154^{**}
	(0.070)
最大值.tp.	0.257^{**}
	(0.115)
标准差.tp.	-0.335^{***}
	(0.123)
最小值（tp）	-0.144^{**}
	(0.058)
最小~最大值（tp）	0.143^{**}
	(0.056)
max.vt.	-0.255
	(0.168)
常数	-0.000
	(0.036)
观测值	348
R^2	0.565
调整后的 R^2	0.537
残差标准误	0.680（自由度 = 326）
F统计量	20.179^{***} （自由度 = 21; 326）
Note:	*p<0.1; **p<0.05; ***p<0.01

（表7）。因此，我们注意到，与 mean.tp. 相比，cv.vt 和 std.vt 的相对排名更高，这与随机森林（表5）和 XGBoost（表6）方法的结果相反。再次，B_z^{ICME}的变化主要由 B_y、n_p和 $|B|$中的特征解释。

我们还可以通过图形来探究这些结果（图2），图中凸显了各变量的相对贡献。特别是，我们注意到，在本例中，前四个变量解释了 B_z^{ICME} 中的大部分变异性。

表5。 用于评估参数相对重要性的随机森林方法。

参数	总体
总体	
max.bx.	-0.2629452
bx 的标准差。	1.3208674
bx 的最小值。	1.4478551
bx 的变异系数。	2.4411125
bx 的极差。	-2.5282641
按均值。	3.3642488
按最大值。	5.4092380
按标准差	8.0053151
按最小值	8.5938032
按变异系数	-1.6325343
按组最小最大值	1.3254579
均值.bz.	1.0951099
最大值.bz.	7.4562721
标准差.bz.	5.8829339
最小值.bz.	3.0405328
变异系数.bz.	0.2845630
极差.bz.	1.9436885
均值	7.1670919
组间最大值	7.7885147
标准差	4.0623384
组间最小值	-2.4643346
bt 变异系数	-0.9527299
minmax.bt	1.6490376
vt.均值	4.3299150
max.vt	4.1455184
vt.标准差	1.3582930
min.vt.	1.7991820
vt.变异系数	1.6273622
minmax.vt.	2.8107001
np.均值	12.5156175
最大 np	5.8775704
std.np.	1.2936570
最小值（np）	11.3617008
变异系数（np）	1.9204835
minmax.np.	2.7836057
mean.tp.	7.5204233
max.tp.	2.9264847
标准差（tp）	3.8501390
最小值（tp）	4.2046862
变异系数（tp）	3.6325805
最小-最大值（tp）	2.9420327

逐步回归法还可以与 Akaike 信息准则（AIC）结合使用，以确定最佳模型，即最能解释输出变量的参数组合。AIC 值越高，越能支持某个变量的重要性。采用这种方法后，按重要性排序，mean.np.、cv.bt. 和 std.bt. 均被发现尤为重要（表8）。

表6. 用于参数相对重要性的 XGBoost 方法。显示前20项。

参数	总体
按最小值	100.000
组间最大值。	88.261
np.均值	32.384
均值	23.530
按标准差	18.147
最小值（np）	17.399
按均值	14.750
mean.tp	13.012
最大 bx。	10.333
np 的最小最大值。	6.411
vt.变异系数	5.980
最大 bz。	5.911
bt 变异系数	5.464
bx 变异系数	5.017
最小-最大值（tp）	4.914
按变异系数	4.499
最大值（tp）	4.311
最小值（tp）	3.853
最小-最大值（bz）	3.809
最小-最大值（bx）	3.754

表7. 用于参数相对重要性的 MARS 方法。结果按 gcv 和 rss 指标所给出的重要性降序排列。

参数	子集数	gcv	rss
按最小值	12	100.0	100.0
np.均值	11	59.8	63.1
均值	10	49.8	53.8
最小值（np）	8	36.1	40.9
vt.变异系数	7	26.9	33.0
vt.标准差	5	18.3	24.7
最大 np	5	18.3	24.7
mean.tp.	4	17.9	23.2
std.np.	4	17.1	22.5

最后，我们考察 DALEX 软件包。严格来说，它是一个元软件包，可以比较不同模型的响应，从而进行直接比较。不过，在这里，我们运用 DALEX 的方法，结合随机森林技术，并使用“explain”和“变量重要性”函数来量化各变量的相对重要性（表9）。在本例中，np.均值、组间最大值、按最小值、按标准差、最小值（np）和均值这六个变量描述了B_z^{ICME}中大部分观测到的变异性；均值与下一个变量 max.np 之间存在明显的差距。这些变量对整体变异的贡献程度也可以通过图形直观呈现（图3）。据此，我们推断，前六个变量涵盖了大多数观测值。

表8. 用于评估参数相对重要性的逐步回归法。仅显示最重要的10个参数，并按重要性从低到高排列。

参数	Df	平方和	RSS	AIC
none	...	...	156.29	-246.56
...	...	...	...	...
变异系数（np）	1	1.8425	158.13	-244.49
最大 bz 值	1	2.6053	158.90	-242.81
标准差 bz.	1	3.6149	159.91	-240.61
均值	1	4.1432	160.44	-239.46
最大 np	1	5.2433	161.53	-237.08
mean.tp.	1	6.4323	162.72	-234.53
按最小值	1	6.4458	162.74	-234.50
np.均值	1	9.2595	165.55	-228.54
bt 变异系数	1	15.5157	171.81	-215.63
标准差	1	15.7451	172.04	-215.16

表9。 DALEX 逐步回归法中参数的相对重要性。

No.	变量	平均随机失活损失	标签
1	完整模型	0.3336585	随机森林
2	目标	0.3336585	随机森林
...	...	...	...
35	标准差	0.3660815	随机森林
36	最大 np	0.3789919	随机森林
37	按最大值	0.3822500	随机森林
38	均值	0.3975788	随机森林
39	按标准差	0.4044210	随机森林
40	最小值（np）	0.4075862	随机森林
41	按最小值	0.4163421	随机森林
42	组间最大值。	0.4368321	随机森林
43	np.均值	0.4562865	随机森林
44	基线	1.2421413	随机森林

3.1 使用更新后的数据库重新分析

尽管我们关注的是 Reiss 等（2021）报告的同一数据集，但我们使用更新后的行星际日冕物质抛射数据集（v2.1）重复了分析，该数据集共包含 521 个事件。我们发现，结果总体上与先前的分析一致，但关联程度总体略差，这可能是因为近期的日冕物质抛射整体较弱。因此，虽然主要解释变量保持不变（例如，由 np 和 bt 得出的变量），但重要性较低（但仍具有统计显著性）的变量在各表中的排序略有调整。

最后，我们重复了这项分析，这次将事件分为此前发生过激波和此前未发生过激波两类。结果总体符合我们的预期：对于包含激波的组，变量之间的关系最强；对于不包含激波的组，关系最弱。不过，这种影响较为有限。例如，调整后的-R^2是一项模型准确度指标，旨在减轻过拟合的影响；在本研究中，它用于处理不同数据集之间的样本量差异。其数值分别为：无激波数据集、所有事件数据集和仅含激波的数据集对应 0.500、0.524 和 0.542。因此，随着数据集中激波所占比例增加，调整后的-R^2 也随之增加，这

符合我们的直觉预期；不过，三个数据集之间的差异也相对有限。

4 总结与讨论

本研究发现，反映数密度和磁场强度的统计量能够解释行星际日冕物质抛射中大尺度磁场 z 分量变异性的很大一部分。综合来看，这些特征反映了行星际日冕物质抛射对其前方背景太阳风的压缩程度。直观而言，这合乎情理：日冕物质抛射喷发时获得的能量会分配到磁能密度和动能中。因此，能量更强的日冕物质抛射会以更强的磁通绳磁场（更大的 B_z）和更高的速度发射。速度更快的日冕物质抛射会造成更强的鞘层压缩，使其前方的数密度和总场强增加。

在某些方法中，还发现了其他可能相关的特征，例如鞘区内的平均质子温度。这也与该区域的压缩有关，因为等离子体受挤压时会升温。温度未被一致地发现对于解释抛射物中的磁场变化同样重要，这可能与这些测量值本身存在的噪声有关，从而削弱了信号成分。还值得指出的是，与 v_x 和 B_x相关的特征在受到压缩时不会发生变化；在大多数结果中，这些特征明显缺失（或重要性较低），进一步支持了这样一种观点：压缩是连接这些特征与目标的关键机制。

在我们的研究中，鞘区是否由激波界定似乎并未造成显著差异。因此，这种方法也适用于强度不足以产生可观测激波的事件。所需的条件只是存在鞘区。另一方面，对于伴有激波的事件，这种关联更强。因此，我们认为这种方法对这类事件可能更准确。此外，将较近期的事件纳入数据集后，关联程度略有降低。这至少初步表明，可能存在太阳活动周效应，不过还需要进一步研究来支持或否定这一观点。

应当强调，这些结果本质上是统计性的。我们始终面临过度解读结果，或将因果关系归因于某个看似重要、实则可能只是偶然出现的特征的风险。与 B_y 相关的那些特征就是一个很好的例子。min.by 有何意义？它似乎不像 mean.by 或 max.by 那样，能以较为直观的方式直接反映压缩。我们可以设想这样一种机制：与正值相比，负 By 值（即较大的最小值）可能更倾向于与某一磁扇区（向内）相关，而非另一个磁扇区；但由于数据覆盖了一个完整的太阳活动周，目前尚不清楚这种不对称性是否存在。此外，即使确实存在，其背后的机制也不明确。它也可能只是对整体磁场压缩程度的间接衡量，而且由于受噪声影响小于例如按最大值的 max.by，因此表现得更明显。由于 min.by 的相对重要性会随所采用的机器学习方法而变化，我们还必须认识到，它的存在可能不过是一种统计异常。正如前文所述，不同技术依赖不同的目标函数，因此，特征的最优优先级排序预计会因技术不同而略有变化；当然，随着排序重要性逐渐降低，这种变化也会更加明显。

我们的结果可以与更传统的方法进行比较，这些方法用于关联太阳风结构的上游与下游属性。例如，Borovsky 和 Denton（2010）对共转相互作用区进行了叠加历元分析，以量化穿越流界面时太阳风参数的变化。最近，Regnault 等人（2020）采用类似技术揭示了 ICME 剖面的一些普遍特征，

他们发现，快速行星际日冕物质抛射在 1 天文单位处仍保留着压缩迹象，这与我们的结果一致。此外，他们还成功重建了磁场的总体非对称剖面，而我们的方法无法做到这一点。最后，Salman 等人（2020）采用统计方法和叠加历元法分析了行星际日冕物质抛射及其相关鞘区的属性。他们将事件分为：（1）没有鞘层的日冕物质抛射；（2）有鞘层但没有激波的日冕物质抛射；（3）同时具有鞘层和激波的日冕物质抛射。他们发现，后两类与日冕物质抛射的速度密切相关。总之，他们发现，驱动激波的鞘层往往压缩程度更高（密度和磁场强度更高），温度也更高。

本研究可为若干后续研究提供思路。例如，利用更新后的 v2.1 日冕物质抛射目录，可使用该数据集预先筛选出可能最具地球有效性的事件。这些事件很可能是速度最快、能驱动最强鞘区的行星际日冕物质抛射。我们预计，这些事件中的关联会更强。此外，还可利用其他对行星际日冕物质抛射鞘层压缩特性的估算方法，例如日球层成像仪或行星际闪烁测量，来估计行星际日冕物质抛射内部磁场的 z 分量可能值；原则上，这有望延长预报提前量。

我们的结果揭示了在包含噪声以及由复杂物理过程形成的信号的数据集中应用复杂统计分析时，其固有的敏感性和局限性。虽然这些方法可能揭示各项特征的真实相对重要性，但有时未必能够将其与数据中的噪声所导致的统计波动区分开来。对数据采用不同的重新归一化方法进行重新分析，进一步凸显了这一点。尽管这并未实质性影响本研究的主要结果，却改变了部分次要但统计显著特征的相对重要性排序。尽管存在这些局限，我们仍认为，对同一数据集应用多种方法是一项有价值的工作，因为这使我们能够评估仅关注单一方法是否会导致对结果的过度解读。

最后，我们重申本研究的主要结论：(1) 可以利用日冕物质抛射上游原位探测航天器易于获取的测量值，通过机器学习技术推断行星际日冕物质抛射内部的磁场 z 分量；(2) 最重要且具有统计显著性的特征与鞘层等离子体密度和磁场强度有关；(3) 对这一结果的物理解释直观易懂：这些参数反映了日冕物质抛射上游的压缩程度，而压缩程度与行星际日冕物质抛射的速度相关，进而也与抛射物内部磁场的整体强度相关。

致谢
PR 谨此感谢美国国家航空航天局（80NSSC18K0100、NNX16AG86G、80NSSC18K1129、80NSSC18K0101、80NSSC20K1285、80NSSC18K1201 和 NNN06AA01C）、美国国家海洋和大气管理局（NA18NWS4680081）以及美国空军（FA9550-15-C-0001）提供的支持。MAR 和 CM 的部分经费由欧盟（欧洲研究委员会，HELIO4CAST，101042188）提供。然而，所表达的观点和意见仅代表作者本人，并不一定反映欧盟或欧洲研究委员会执行局的立场。欧盟和资助机构均不对此承担责任。”

开放研究

数据可用性声明

本研究使用的所有数据和代码均可从 Zenodo 获取（Riley, 2023）。

参考文献

Bailey, R. L., Reiss, M., Arge, C. N., M¨ostl, C., Owens, M. J., Amerstorfer, U., . . . Hinterreiter, J. (2020). 利用机器学习改进背景太阳风模型预测。*arXiv* 预印本 *arXiv:2006.12835*。Bendel, R. B., & Afifi, A. (1976). 一种逐步回归准则。《美国统计学家》, *30*(2), 85–87. Biecek, P. (2018). Dalex：用于复杂预测模型的解释工具。《机器学习研究杂志》, *19*(1), 3245–3249. Bobra, M. G., & Couvidat, S. (2014). 使用机器学习算法和 SDO/HMI 矢量磁场数据预测太阳耀斑。*arXiv* 预印本 *arXiv:1411.1405* 。Borovsky, J. E., & Denton, M. H. (2010). 太阳风湍流与剪切：对 1 天文单位处共转相互作用区的叠加历元分析。《地球物理研究杂志：空间物理》, *115*(A10). Camporeale, E. (2020). ML-Helio：日球物理学与机器学习交叉领域的新兴社群。《地球物理研究杂志：空间物理》, *125*(2). Camporeale, E., Care, A., & Borovsky, J. E. (2017). 利用机器学习进行太阳风分类。《地球物理研究杂志：空间物理》, *122*(11). Camporeale, E., Wing, S., Johnson, J., Jackman, C., & McGranaghan, R. (2018). 机器学习时代的空间天气：一种多学科方法。*Space Weather*。Chen, J., Cargill, P. J., & Palmadesso, P. J. (1996). 地效太阳风结构的实时识别与预测。《地球物理研究快报》,*23*(6), 625–628. Chen, T., & Guestrin, C. (2016). XGBoost：一种可扩展的树提升系统。载于第*22*届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集 (pp. 785–794)。Gosling, J. T., & McComas, D. J. (1987, April). 快速日冕物质抛射周围的磁力线绕流——强烈日球面外行星际磁场的来源。《地球物理研究快报》, *14*, 355-358. doi: 10.1029/GL014i004p00355 Gr¨omping, U., et al. (2006). R语言线性回归的相对重要性：relaimpo 程序包。《统计软件杂志》, *17*(1), 1–27. Heidrich-Meisner, V., & Wimmer-Schweingruber, R. F. (2018). 使用 k 均值聚类算法进行太阳风分类。载于《空间天气机器学习技术》(pp. 397–424)。爱思唯尔。Horbury, T. S., & Balogh, A. (2001, August). 高速太阳风流中磁场波动的演变：尤利西斯号和 Helios 的观测。《地球物理研究杂志》, *106*, 15929-15940. doi: 10.1029/2000JA000108 James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). 《统计学习导论》(Vol. 112)。Springer。Liaw, A., Wiener, M., et al. (2002). 使用随机森林进行分类与回归。《*R*新闻》, *2*(3), 18–22. Lugaz, N., Farrugia, C. J., Winslow, R. M., Al-Haddad, N., Kilpua, E. K. J., & Ri-ley, P. (2016, November). 影响 1 AU 处激波和鞘层地效性的因素。《地球物理研究杂志》, *121*, 10. doi: 10.1002/2016JA023100 M¨ostl, C., Bailey, R. L., Amerstorfer, U. V., Amerstorfer, T., Weiss, A. J., Reiss, M. A., . . . Bauer, M. (2020). HELIO4CAST：用于提升地球空间天气预测能力的实时测试环境。载于欧洲地球科学联盟大会摘要 (p. 5202)。M¨ostl, C., Isavnin, A., Boakes, P. D., Kilpua, E. K. J., Davies, J. A., Harrison, R. A., . . . Zhang, T. L. (2017, July). 利用日球层成像仪对太阳日冕物质抛射进行建模，并通过日球物理系统观测台验证。*Space Weather*, *15*, 955-970. doi: 10.1002/2017SW001614

Nieves-Chinchilla, T., Vourlidas, A., Raymond, J. C., Linton, M., Al-Haddad, N., Savani, N., . . . Hidalgo, M. A. (2018). 利用 20 多年的 Wind观测了解行星际日冕物质抛射的内部磁场构型。《太阳物理学》, *293*(2), 1–31. Qahwaji, R., Colak, T., Al-Omari, M., & Ipson, S. (2008). 利用机器学习和日冕物质抛射—耀斑关联自动预测日冕物质抛射。《太阳物理学》, *248*(2), 471–483. R Core Team. (2019). R：用于统计计算的语言和环境[计算机软件手册]。维也纳，奥地利。取自 https://www.R -project.org/ Regnault, F., Janvier, M., D´emoulin, P., Auch`ere, F., Strugarek, A., Dasso, S., & No^us, C. (2020). ACE数据20年：叠加历元分析如何揭示行星际日冕物质抛射剖面的普遍特征。《地球物理研究杂志：空间物理》, *125*(11), e2020JA028150. Reiss, M. A., M¨ostl, C., Bailey, R. L., R¨udisser, H. T., Amerstorfer, U. V., Amer- storfer, T., . . . Windisch, A. (2021). 利用太阳日冕物质抛射的上游原位观测，通过机器学习预测 Bz磁场分量。*Space Weather*, e2021SW002859. Riley, P. (2023). *Zenodo*。 取自 https://www.zenodo.org/ doi: 10.5281/ zenodo.7730719 Riley, P., Ben-Nun, M., Linker, J. A., Owens, M. J., & Horbury, T. (2017). 利用简单的模式识别预测太阳风特性。*Space Weather*, *15*(3), 526–540. Riley, P., Lionello, R., Miki´c, Z., Linker, J., Clark, E., Lin, J., & Ko, Y.-K. (2007, January). 太阳爆发后发生的“爆发式”重联：磁流体动力学模拟及与观测的比较。《天体物理学杂志》, *655*, 591-597. doi: 10.1086/509913 Salman, T. M., Lugaz, N., Farrugia, C. J., Winslow, R. M., Jian, L. K., & Galvin, A. B. (2020). 基于 STEREO 原位观测研究 1 天文单位附近有无激波的日冕物质抛射鞘区特性。《天体物理学杂志》, *904*(2), 177. Salman, T. M., Lugaz, N., Winslow, R. M., Farrugia, C. J., Jian, L. K., & Galvin, A. B. (2021). 日冕物质抛射驱动的鞘区分类：STEREO事件的特征。《天体物理学杂志》, *921*(1), 57. Savani, N., Vourlidas, A., Szabo, A., Mays, M., Richardson, I., Thompson, B., . . . Nieves-Chinchilla, T. (2015). 预测抵达地球的日冕物质抛射内部的磁矢量：1. 初始结构。*Space Weather*, *13*(6), 374–385. Shiota, D., & Kataoka, R. (2016). 利用内部磁通量绳（Susanoo-CME）对多个日冕物质抛射的行星际传播进行磁流体动力学模拟。*Space Weather*, *14*(2), 56–75. Siscoe, G., & Odstrcil, D. (2008). 行星际日冕物质抛射鞘区与磁层鞘的差异。《地球物理研究杂志：空间物理》, *113*(A9). T¨or¨ok, T., Downs, C., Linker, J. A., Lionello, R., Titov, V. S., Miki´c, Z., . . . Wijaya, J. (2018). 2000年7月14日“巴士底日”爆发的日地磁流体动力学模拟。《天体物理学杂志》, *856*(1), 75. Venables, W. N., & Ripley, B. D. (2002). 现代应用统计学：基于 *S* 的统计计算 （第 4 版）。纽约：Springer。取自 http://www.stats.ox.ac.uk/pub/ MASS4 (ISBN 0-387-95457-0) Witten, I. H., & Frank, E. (2005). 数据挖掘：实用机器学习工具与技术。Morgan Kaufmann。wrapper, S. M. D. f. m. b. T. H. a. R. T. U. A. M. F. u. w. T. L. l. (2019, Novem- ber). *earth*：多元自适应回归样条。取自 2020年6月1日, https://CRAN.R-project.org/package=earth Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). 一种采用提升决策树的方法

用于信用评分的贝叶斯超参数优化。*Expert Systems with Applications*，*78*，225–241。

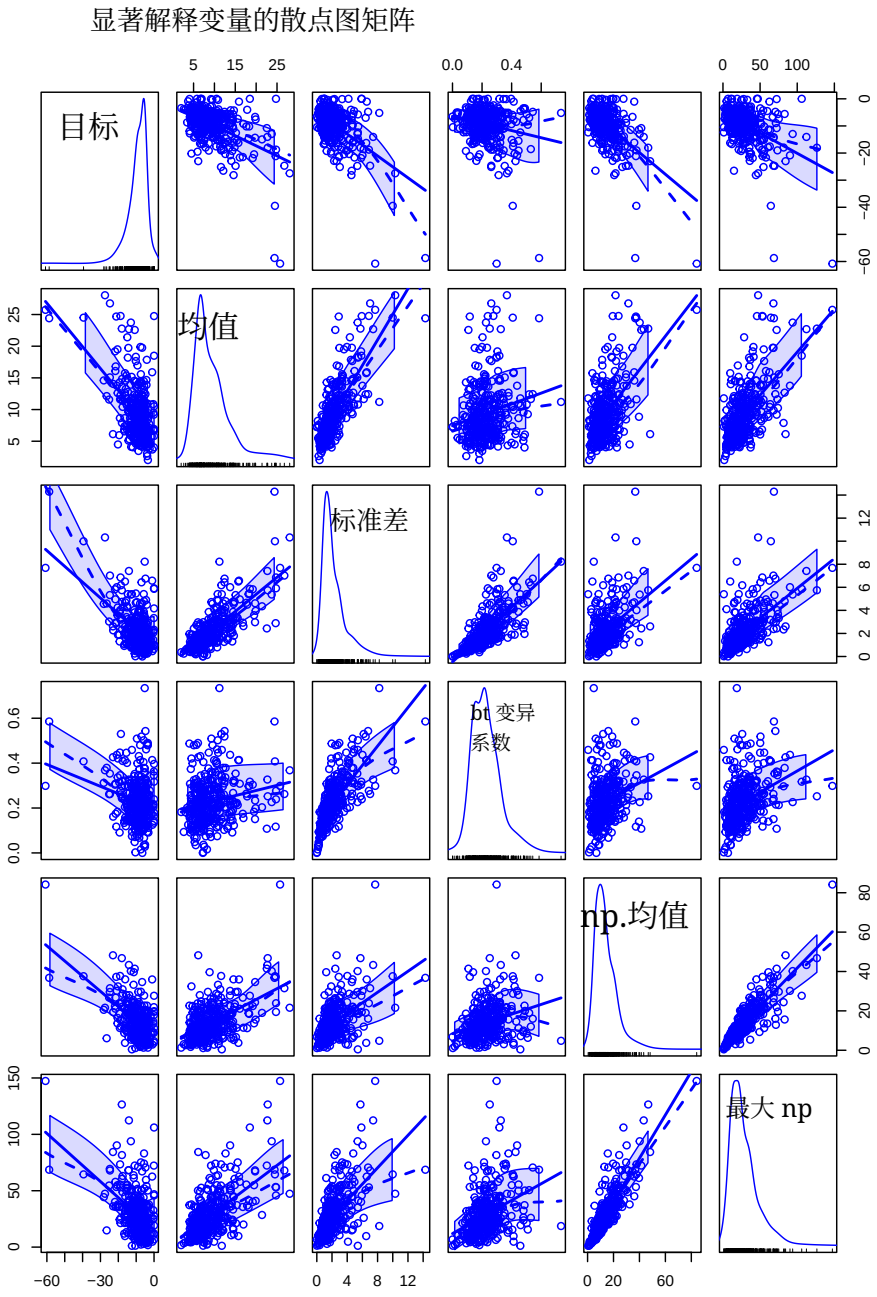


图 1. 行星际日冕物质抛射中 B_z 最小值的五个显著解释变量散点图矩阵。 各面板显示： B_z （行星际日冕物质抛射中的输出或“目标”变量）、均值、组间标准差、bt 变异系数、np.均值和最大 np。圆圈表示数据，实线表示回归线，虚线表示平滑均值，点划线表示方差。

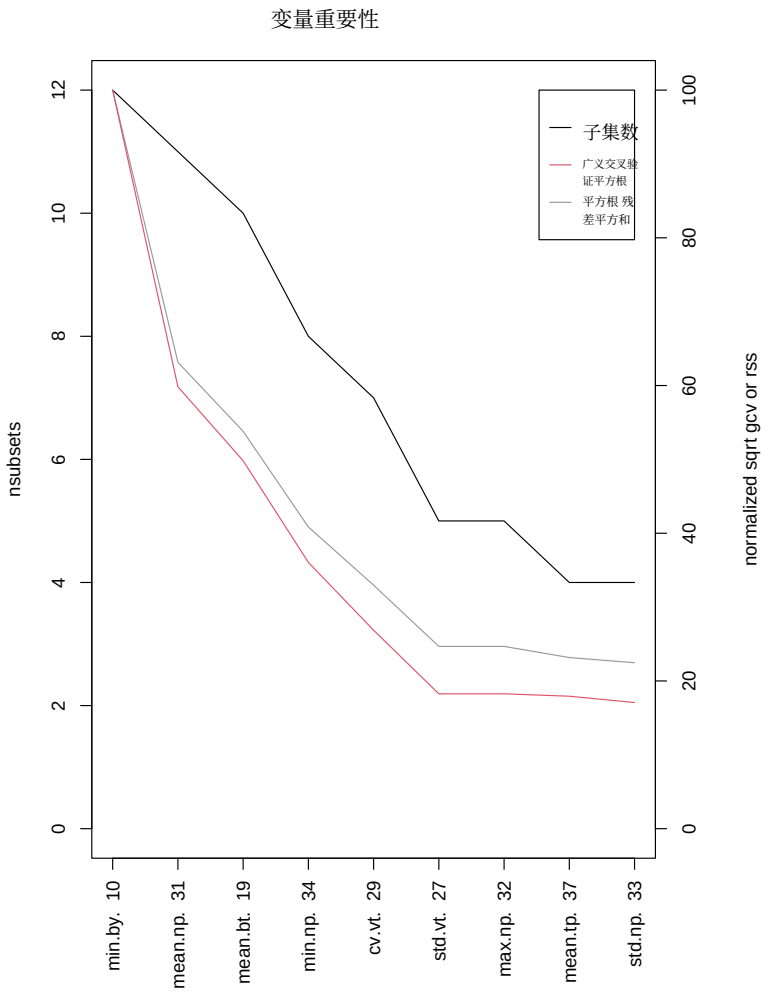


图2。参数相对重要性的 MARS 方法。更多详情请参见正文。

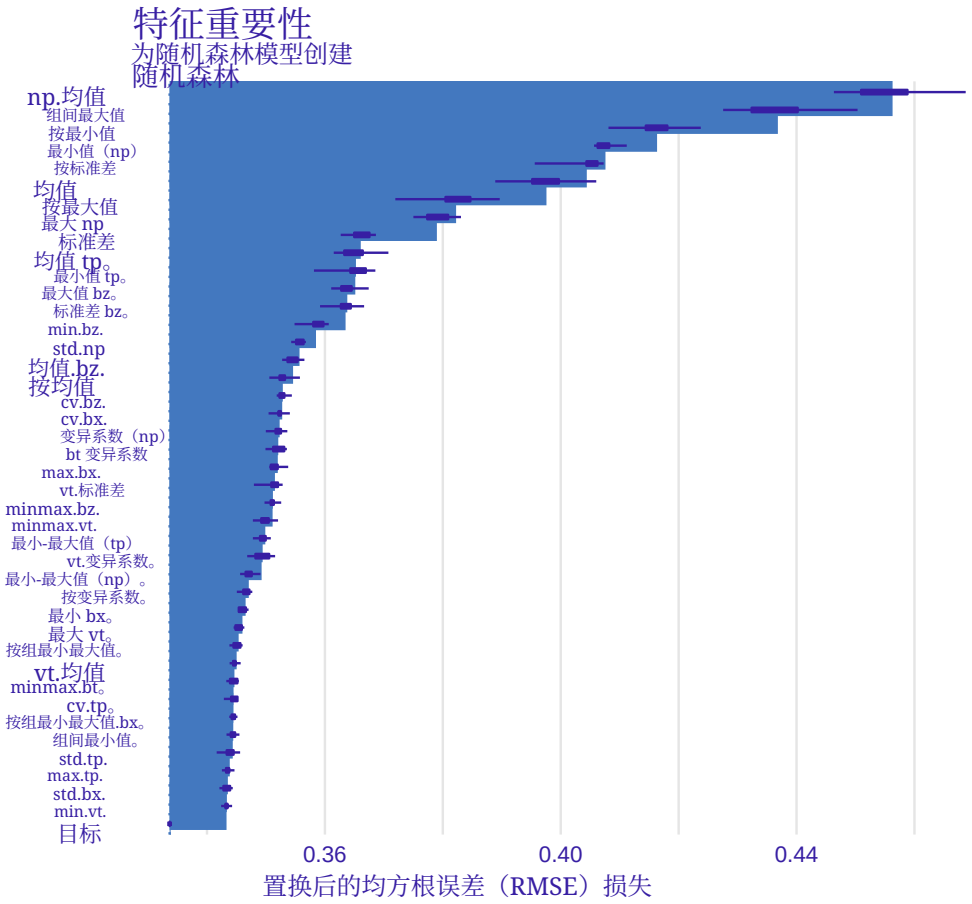


图3. 使用 DALEX方法计算的解释变量相对重要性。RMSE损失是衡量特定输入变量相对重要性的一项指标。