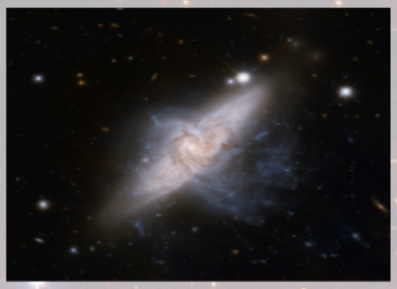
Science Ascend

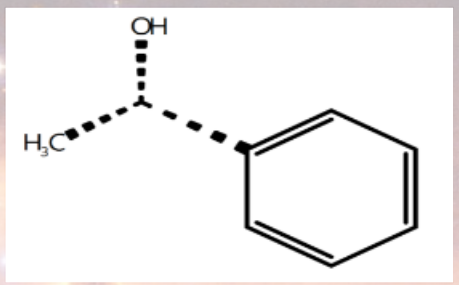
From September 02-06, 2024!

Rising to new heights of discovery with Science!
Every week!



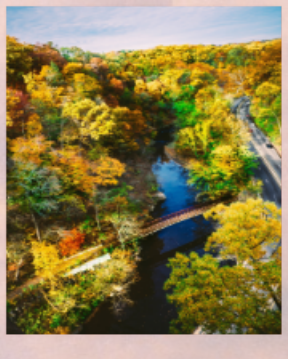
More precise methods,
Higher resolution outputs, ALMA, Keck, GAIA and more!

Faster, better, stronger biosensor, efficient MS and more!



Land use classification *on space!* More robust deforestation surveillance and more!

Low toxic end product degradation of chemicals, NA handling with Bayesian and more!



cout<<solutions;

Stronger theoretical base, robust algorithms, shorter runtime!

ISSN: 3062-0090
FIRE Araştırma Eğitim Ltd. Şti., Vol:1, Issue:1

Science Ascend

Rising to New Heights of Discovery!

Science Ascend teleports you to the frontiers of science. It compiles and discuss the scientific research preprints from arXiv, bioRxiv, chemRxiv just from the previous week to be cognizant of the *state-of-the-art* of knowledge in astrophysics, chemistry, environmental chemistry, remote sensing, and applied statistics/data science. Light from the *Science Ascend* will keep brightening the dark horizon beyond the limits of our comprehension. FIRE Araştırma Eğitim Ltd. Şti. guarantees the weekly publication and dissemination of this journal, and make it available for everyone at most fifteen days after its publication freely.

Publisher: FIRE Araştırma Eğitim Ltd. Şti.
Media: Online Journal
Responsible person: Yasin Güray Hatipoğlu
Editor-in-chief: Yasin Güray Hatipoğlu
Editor: Yasin Güray Hatipoğlu
Authors: Yasin Güray Hatipoğlu
Frequency: Once a week
Address: Yıldızevler Mah. Kişinev Cad. No:10 Çankaya/Ankara/Türkiye
Website: https://fire-ae.github.io
ISSN: 3062-0090

This issue: September 9, 2024

Volume: 1

Issue Number: 1

Bilim Yükselişi

Keşfin Yeni Yüksekliklerine Ulaşmak!

Science Ascend sizi bilimin sınırlarına ışınlar. Astrofizik, kimya, çevre kimyası, uzaktan algılama ve uygulamalı istatistik/veri bilimi alanlarındaki bilgi birikiminin *en son durumu* hakkında bilgi sahibi olmak için arXiv, bioRxiv, chemRxiv'den sadece bir önceki haftaya ait bilimsel araştırma ön baskılarını derler ve tartışır. *Bilim Yükselişi*'nden gelen ışık, kavrayışımızın sınırlarının ötesindeki karanlık ufku aydınlatmaya devam edecektir. FIRE Araştırma Eğitim Ltd. Şti. bu derginin haftalık olarak yayınlanmasını, dağıtılmasını ve yayınlandıktan en geç on beş gün sonra ücretsiz olarak herkesin erişimine açılmasını garanti eder.

Yayıncı: FIRE Araştırma Eğitim Ltd. Şti.
Ortam: Online Journal
Sorumlu Kişi: Yasin Güray Hatipoğlu
Yazı İşleri Müdürü: Yasin Güray Hatipoğlu
Editör: Yasin Güray Hatipoğlu
Yazarlar: Yasin Güray Hatipoğlu
Yayınlanma Sıklığı: Haftada bir kez
Adres: Yıldızevler Mah. Kişinev Cad. No:10 Çankaya/Ankara/Türkiye
Website: https://fire-ae.github.io
ISSN: 3062-0090

Bu sayı: September 9, 2024

Cilt: 1

Sayı Numarası: 1

Last week in Astrophysics

Author: Yasin Güray Hatipoğlu

This section compiles and discusses arXiv preprints from the September 2-6, 2024 week from astro.EP category. These preprints are not peer-reviewed themselves, but generally a part of the peer-review process of another full-research article, or done with rigor and from researchers in these fields. Hence, with a curious and critical mind, readers can scrutinize them to see the cutting-edge, instead of at least months-behind research articles.

Bakx and Conway [1] emphasized the need to properly cite the instrumentation that a research paper utilized in its data. Their focus was on the Atacama Large Millimetre/submillimetre Array (ALMA). They created a survey in the European Southern Observatory workshop in June 2024 and provided papers to cite for different parts and bands of the instrument. The main drive was that proper acknowledgment would boost higher-performing instruments and this will attract later funds to improve the research in these areas.

Wang et al. [2] investigated the source of 1/f noise[1] in the heliosphere. Authors considered that working through this 1/f noise and examining related scale-invariant signal-generation physical processes will further our understanding about heliosphere, too. They especially consider NASA's Polarimeter to Unify the Corona and Heliosphere (PUNCH) mission data. Several previous approaches involved the assumption of superposition of several yet-unknown origin processes, and others assigned the solar wind 1/f signal to specific local parts or solar/coronal-related processes. Of course, to probe lower and lower frequencies of this signal, researchers require longer and continuous data segments as well. The authors also stated that providing an accurate interpretation, PUNCH images will have the necessary scales and fine resolution to probe this 1/f signal.

Colzi et al. [3] prepared a preprint on galaxy-scale astrochemistry where they report specific projects examining interstellar medium (ISM)[2] chemistry. We can first list the specific relevant projects like this: ACES Large Program, GOTHAM, QUIJOTE, G31 Unbiased ALMA sPectral Observational Survey-GUAPOS, ALMA Evolutionary study of High Mass Protocluster Formation in the Galaxy-ALMAGAL. They studied this galactic astrochemistry in three different chapters as chemical complexity in the galaxy, isotopic ratios in the galaxy, and the importance of cosmic rays. Starting from the cosmic rays, a constant cosmic ionization rate was found inaccurate in several studies, and even a map of this value was generated. This rate was estimated to be higher in protostellar surfaces and related shocks, and also varying in star-forming regions in general. Hence, a must-consider variable, not a parameter for astrochemical models that involve cosmic ray effects. Secondly, isotopic ratios are related to both stellar evolution (nucleosynthesis processes) and galactic evolution, and different atomic species have different stories through these journeys. Special discussions were provided on Nitrogen 14/15 and Carbon 12/13 isotopes, and the topic ended with the ALMAGAL survey's outputs pointing toward local fractionation processes, and optical depth. Authors recommended a joint study involving single-dish and interferometric[3] observations together. On chemical complexity in the galaxy, authors reviewed a galactic center molecular cloud G+0.693-0.027 and a galactic disk star-forming region G31.41+0.31 hot molecular core. Judging from the linewidths of the spectrum, HII regions, H_2O masers[4], dust continuum point sources, cloud-cloud collision velocity shocks and many atomic and molecular species compositions.

Buana et al. examined a more accurate way to classify the hazard status of asteroids via machine learning and deep learning technics[4]. It is an empirical approach where a compendium of ML and DL methods are being applied to the labeled datasets and a standardized score is used to test the performance of these methods. They used a Kaggle dataset (958.524 samples, 45 features) and also mined the Near Earth Object Website-NeoWS platform of NASA (362717 samples, 41 features). They utilized 5 ML (logistic regression, k-nearest neighbor, support vector machine, random forest, and catboost) algorithms, and 5 DL (multilayer perceptron, deep neural network, convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory RNN)[5]. Firstly, they indicated the imbalance in

[1] A frequently observed type of noise in many different domains, higher in the low-frequency part of the power spectrum and gets lower and lower with increasing frequency, a.k.a. pink noise or flicker noise. A critical property of 1/f noise is that integrated power per octave [halving or doubling the frequency] across different frequencies stays the same, and when both frequency and power were expressed in logarithms, the graph resembles 1/x, where x is log(f)

[2] Not within a star or solar system, and not outside of our galaxy. In other words, the medium among the star systems of the Milky Way, excluding the star systems themselves.

[3] Single-dish is one telescope with a main-beam and mirror, while interferometry involves multiple antennas and utilizes interferometry phenomenon, also can be seen in virtual antennas of RADAR earth observation satellites

[4] microwave or molecular amplification by stimulated emission of radiation, similar to laser abbreviation

[5] Authors provided short texts for each method, but

the dataset, as for the Kaggle data, only 0.002 % of the sample is hazardous, while for NeoWS, it is still only 13 %, hence they bootstrapped and synthetic minority oversampling technique (SMOTE) the data to make it evenly distributed, and also removed mostly-NA filled columns. In the end, both in terms of training time required in their computer and accuracies in both datasets, Random Forest outperformed other classifiers.

Zhang et al. [5] reported a discovery of a near earth-mass planet (with 1.9 ± 0.2 earth mass in around 2.1 ± 0.2 astronomical unit (au) and a brown dwarf orbiting a white dwarf, an end product after red-giant phase. They worked on KMT-2020-BLG-0414, an ultra-high-magnification microlensing event to examine. The crucial thing is that this potential earth-mass planet survived the engulfment by its star in the red-giant phase. They used Keck Observatory's NIRC2 camera data and provided the reduced (processed) data in a Zenodo link, they also used VISTA Variables in the Via Lactea Survey (VVV) survey data and Open Gravitational Lensing Experiment (OGLE) data. To begin with, the star they were focusing on was either a main sequence star or a white dwarf, and it was only possible to ensure which one was the actual case via pre-event and/or post-event photometry with a proper angular-resolution observatory. Briefly, the authors concluded that it is much more likely to be a white dwarf after observing with K-short infrared passband. Four different models to explain observed lightcurve data were constrained with these additional data. They also further constrained the potential orbital model, specifically physical separation, semi-major axis from projected separation with Bayesian inference. They concluded that a substantially favored orbit is the planet in the near greatest elongation with a significantly inclined orbit. Another not ruled-out orbit is in near conjunction on a nearly edge-on orbit case.

In another study utilizing ALMA data, Speedie et al.[6] checked the evidence for gravitational instability in planet formation. Normally, the widely-held theory for planet formation is *core accretion*. According to this model, roughly, very small bodies keep coalescing within the disk, and again many of these coalesced materials sink to their respective star. However, at some point, rather than sinking to the star, the accreted material keeps accreting new masses and instead starts to clear its debris. In gravitational instability, though, especially for giant planets and at least 1/10 mass ratio between the circumstellar disk and the star, the planet forms comparatively rapidly. The authors aimed for AB Aurigae disk, classified as a Class II Young Stellar Object. Their data was ALMA Band 6 observations of ^{13}CO (J=2-1) and $C^{18}O$ (J=2-1) in channel widths of 42, and 84 m/s, respectively. They created image cubes and then converted them to 2D-moment maps of velocity-integrated intensity, intensity-weighted line-of-sight velocity, and emission line width. Spiral arms in the disk were revealed by high-pass filtering the ^{13}CO moment maps. They also make a 3D smoothed-particle hydrodynamic simulation of gravitationally unstable disk with PHANTOM code, aiming to find a qualitative comparison to the actual data. Later, relevant to the gravitational instability thermal gradients, they utilized the Monte Carlo radiative transfer code MCFOST. Later, the giggle package was used to compute velocity fields of gravitationally unstable disks. They provided all data, reduced data, codes, and the codes they used with links. In the end, they consider this as a gravitational-instability-driven planet formation, with reported compatible values from simulations regarding disk mass ratio and cooling timescales.

Evans et al. [7] examined triple-star system planetary orbits and reported that their 21 stellar systems had at most one plane of alignment instead of coplanarities. From the Kepler observatory, they chose observations done with laser guide star AO system, with mostly *K'* filter but also K_{cont} filter when *K'* would likely be saturated from the brightness of the target, and also non-redundant aperture masking interferograms. For most samples, they also obtained Hobby-Eberly Telescope's red arm of the second-generation Low-Resolution Spectrograph (LRS2-R) red-optical spectra at McDonald Observatory. After assembling this telescope's data, with unresolved broadband photometry of r' i' z' J H K_s from Kepler Input Catalog, 2-Micron All-Sky survey - 2MASS, and Keck NIRC2 high-resolution adaptive optics imaging. Custom modified Gibbs algorithm and emcee fitted to check stellar parameters MESA Isochrones and Stellar Tracks (MIST) evolutionary models generated prior for stellar radii, and when available, Gaia Data Release 3 provided prior for parallax. After stellar and planetary parameter revisions, orbital arcs and orbital alignment were measured and tested as well. These arcs were compared to simulated orbit arcs, and also planet-hosting binaries. Finally, they made a full orbital analysis, for inner binaries using ORVARA, and for outer companions using LOFTI-GAIA. The authors plan to improve the precision of orbital characteristics by further monitoring nine compact triple systems.

Coria et al. [8] studied WASP-77A b formation and migration by checking the stellar system's carbon-oxygen isotopic abundances inven-

FIRE suggest this Machine Learning for Intelligent Systems course materials from Cornell University

tory. For the stellar atmosphere, MARCS stellar atmosphere models and local thermodynamic equilibrium (LTE) radiative transfer code TurboSpectrum were utilized. For line databases, Vienna Atomic Line Database (VALD), Kurucz Atomic and Molecular Database, and high-resolution transmission molecular absorption database (HITRAN) were utilized. The ratio of carbon 12 to 13 was derived with their statistical and systematic uncertainties together from ^{13}CH spectral lines and found 66 $\pm$ 18 from HIRES, and 51 $\pm$ 6 from ESPRESSO spectrum analysis for WASP-77 A, and for its planet WASP-77A b this ratio is 26.4 $\pm$ 16.2, comparatively enriched in ^{13}C. Finally, WASP-77A b was considered to be formed beyond the H_2O and CO_2 lines and then migrated to its now approximately 0.024 AU distance from its star.

Föhring et al. [9] studied locations for the Second Flyeye Telescope on Earth to complement the first Flyeye and both before and after Vera Rubin Observatory (or Large Aperture Synoptic Survey Telescope-LSST). They considered three different locations, 1-) right beside Flyeye-1 in Mufara, Italy, 2-) southern hemisphere, La Silla, Chile approx. the same longitude as LSST, and 3) Sutherland, South Africa with 6 hours of separation from LSST. Even though they could not find major statistical outputs in favor of a specific location, the highest number of clear nights in La Silla, and more discoveries of large objects when Flyeye pair is one in north and one in south, favored La Silla for the site selection.

Goswamy et al.[10] made a cross-check study on the Wide Angle Search for Planets (WASP) survey's comparatively-dense hot Jupiters. Authors suspected that this anomaly in density estimations might have arisen from an undetected stellar binary, which can contaminate planet mass and radius estimation, and result in denser-than-actual density estimations. Precise parallax measurements from Gaia Data Releases also permitted stellar angular diameter estimation from the Infrared Flux method (IRFM), using effective temperature and bolometric flux received at Earth. They then compared this to transit profile photodynamical fits. Authors estimated angular diameters by fitting apparent magnitudes in Gaia BP, G, and RP, 2MASS J, H, and Ks, and Wide-field Infrared Survey Explorer (WISE) W1 and W2, with synthetic photometry from stellar model atmospheres of Castelli & Kurucs (2003)[11]. TEPCat is used to extract WASP system data. IRFM procedure was done with in-house developed Python routines with astroquery, pysynphot, and pyphot packages. After their analysis, they identified eight outliers with a high ratio of IRFM radius to that of photodynamical one from transit. They elaborated on seven cases and on average, the expected density of host-star was estimated to be reduced by 1.3 times. The one not elaborated, WASP 86, was already been in rectification of their parameters with additional studies.

Brown and Ogilvie[12] pointed out the issues that arose from 2D models of planet-disc interactions on verticular structure dynamics. Especially considering current calculations in migrations of objects within the disc, the core accretion model would inwardly migrate the object towards the star. Especially the softening length parameter, to mimic 3D characteristics in a 2D model, has a wide range of reasonable values and this variation is a non-negligible issue. Using their approach they found strong agreements with 3D numerical simulations. They were able to capture spiral wakes and horseshoe streamlines simultaneously.

Laurent-Varin et al.[13] utilized Rosetta's both Doppler and optical data (as opposed to previous Doppler data) to work on the gravity field of comet 67P/C-G. They used French Centre National d'Etude Spatiales (CNES) GINS and DYNAMO softwares. They first conducted a gravity sensitivity analysis to check the potential observable values by Rosetta. They had two separate scenarios for parameter estimation, and they also removed "bad" measurements systematically before this parameter estimation in GINS software. They reported that the co-utilization of optical measurements with Doppler measurements increased the number of converged arc measurements significantly.

Arakawa et al.[14] made a theoretical study on oxygen isotope exchange between dust aggregate and ambient vapor to better understand spatiotemporal evolution of solar nebula. They divided this phenomenon into four steps 1) supply of gas molecules to the surface of the aggregate, 2) diffusion of gas molecules within the aggregate, 3) isotope exchange on the surface of aggregate, and lastly 4) isotope diffusion within the particles. They showed that 2) is orders of magnitude faster and cannot be the rate-limiting step, for crystalline cases, the exchange process is always slower than all other processes in all temperatures, and particle radius determines whether the supply of gas to the surface of aggregates or supply of molecules at the surface of particles, with higher radius time for surface of aggregate is larger.

Meier et al.[15] worked on moon-forming giant impacts to Earth[6] considering rotating bodies. They used Smoothed Particle Hydrodynamics code Gasoline and analyzed the post-impact state

[6] *Giant impact hypothesis:* an object hits the ancient Earth and a debris disk appeared after this collision aggregated into Moon in time

with SKID. They simulated 7649 collisions and analyzed 6247 of them, the remaining had insufficient resulting earth mass, or hit-and-run collisions, or unresolved graze-merge cases. Free parameters were rotational configuration, total initial angular momentum, asymptotic relative velocity (pre-impact), impactor-to-target mass ratio and angular velocity factor. The post-impact phase was fixed to seven days. Both objects had three different spin orientations composed of upward, non-rotating, and downward, resulting in nine different categories to consider in simulations. Pre- and post-impact momentum budgets (strong correlation as little mass was lost with negligible momentum), parameters, and hot-spin stability limit. Even though they have promising cases, not all constraints are still satisfied, and no single simulation managed to reach the set constraints. Some of the most promising collisions required very similar compositional similarities between the target and the impactor.

Lovis et al.[16] presented RISTRETTO, a visiting instrument for visible high-resolution spectrograph to be used on ESO Very Large Telescope (VLT) for reflected light exoplanet spectroscopy, especially the nearby ones like Proxima Centauri b.

Table 1: RISTRETTO characteristics

Inner Working Angle	< 40 mas
Planet Coupling Efficiency	$> 50\%$
Stellar Coupling (contrast)	$< 10^{-4}$
Spectral Resolution	$> 100,000$
Spectral Range $\Delta\Lambda/\Lambda$	25%
Total System Throughput	$> 5\%$

In addition to the list of nearby exoplanets, other science cases are expected to be accreting protoplanets in the H-alpha region, kinematics of protoplanetary disks, spatially resolved stellar surfaces, and solar system science.

Last week in Chemistry

Author: Yasin Güray Hatipoğlu

This section lists preprints from the September 2-6, 2024 period from the chemRxiv Analytical Chemistry category.

Adib et al.[17] worked on *in situ* pH-controlled electrochemical biosensors[7] for glucose and pH detection in saliva medium. They preferred interdigitated array[8], biosensing element was Glucose Oxidase (GOx) and glucose detection was successful in 0.02 to 7 mM range (limit of detection[9] is 0.3 μM), while for pH detection it was 5 to 9. They characterized the electrodes with optical microscopy, electrochemical impedance spectroscopy (EIS), scanning electron microscopy (SEM), energy dispersive X-ray spectroscopy (EDX), and atomic force microscopy (AFM). They utilized these examinations to monitor surface topography changes after electrode modifications. The sensor lifetime was around 70 days.

Esselman et al.[18] constructed a workflow to streamline region-of-interest imaging by Matrix-assisted laser desorption/ionization imaging mass spectrometry (MALDI-IMS)[10] to increase efficiency and reduce required time and resources. Their three case studies were on manual annotation of human brain tissue, automated segmentation of renal functional tissue units with custom segmentation algorithms, and automated segmentation of HeLa cells[11] In all cases, the developed workflow greatly enhanced the throughput and focused on regions of interest with much less required resources. HeLa cells were detected automatically by the QuPath Cell Detection feature. For glomeruli in the kidney, an in-house machine learning method developed elsewhere (a deep learning methodology - Mask-RCNN (101 ResNet backbone) from detectron2 package[19]) was used.

Fuku and Yoshida[20] conduct an unsupervised image classification after converting infrared transmission spectra to vector images. They utilized the spectral database (SDBS) of the National Institute of Advanced Industrial and Science Technology and hierarchically clustered these grayscale images. They used the Anaconda platform and OpenCV for image handling, shared data (password protected, require correspondence with Kentaro Fuku), and methodology in GitHub here. The hierarchical clustering divided the 227 compounds into seven main categories. They interpreted all these categories and cautioned that spectral deviations are highly influential in clustering results.

[7] Biosensor is, as it name suggests use biological means to sense something. Enzymes are very specific for certain compounds, and in these electrochemical biosensors, when a substrate, here glucose, binds to the enzyme, a voltage difference emerges and lets us find the glucose.

[8] comparatively higher surface area and better sensitivity, with several other benefits

[9] Limit of detection is usually 3 times of standard deviation while measuring a blank solution using the developed method.

[10] The sample is on a 2-dimensional plate, one desorps a specific part with laser and analyze the desorbed material in MS, and in this way one can create a 2-dimensional map of chemical species as well.

[11] cancerous cells of HEnrietta LAcks, one of the most used matrices in scientific research.

Last week in Remote Sensing

Author: Yasin Güray Hatipoğlu

These preprints are from arXiv's advanced search filtering the September 2-6, 2024, and the words "remote sensing" in all fields.

Le et al.[21] studied on-air land use classification methods for satellite remote sensing. Ideally, the computationally efficient, noise-robust on-air model should be developed for this task. A pre-trained model, specific for a certain task, in this case, remote sensing image classification for land use identification was deemed the best choice by the authors of this study. They compared convolutional neural network-based, residual network-based, and transformer-based models with Vision Transformer (ViT). The dataset was the public benchmark for land use and land cover classification, EuroSat, from Sentinel-2[12] satellite imagery (13 bands, 64x64 pixel size 27000 georeferenced images). Especially their model EfficientViT-M2 handled noisy conditions quite well and was considered the optimum choice for this task. They also applied Gaussian noise and motion blur from the instrument to check the noisy-case performance of these algorithms. They compared training and inference times and confusion matrix results of all approaches. Care should also be taken as some of the algorithms' input size for images was substantially different. Whereas classical ResNet, CNN, Compact CNN-transformer, and SmallViT ones were using 64x64, for other transformers it was 224x224 and more.

Ray and Skurijhin[22] created a workflow for deep clustering of remote sensing images in three steps, that performed better than zero-shot methods. They first fine-tuned a pre-trained deep neural network DINOv2 on a labeled source remote sensing imagery, then extracted a feature vector from each image in the target set, made a manifold projection into a lower dimensional Euclidean space, and clustering these with a Bayesian nonparametric technique for simultaneous cluster number and memberships. The pre-trained model was obtained from PyTorch Hub. A finetuning step made use of the NWPU-RESISC45 dataset for its apparent generalization. Uniform Manifold Approximation and Projection (UMAP) realized the manifold projection on the grounds that its parameters did not have a strong influence on clustering performance and scalability of very large datasets was not an issue. Clustering used the Dirichlet process as prior and clusters were selected to be in the Gaussian mixture model, hence the DP-GMM for the clustering step with variational inferece using Sci-kit package. Excluding GID and EuroSAT datasets, the developed algorithm performed quite well and generally was better than OpenCLIP and RemoteCLIP.

Borlido et al.[23] examined the optimum superpixel method to detect deforestation using satellite imagery. They used seven channels of Landsat-8 satellite images from a total of nine images. The region was Xingu River Basin and PRODES by the United States Geological Survey provided the ground truth on forest and non-forest classification. They used RGB channels of Landsat, three principal components extracted by Principal Component Analysis, and the best Univariate Marginal Distribution Algorithm (UMDA) Individual composed of B4, B3, B1, and B6 bands of Landsat-8. Ground truth was at its smallest 70 pixels, so superpixels were ensured to be at least this large. They considered boundary recall, under segmentation error, explained variation, similarity between image and reconstruction from superpixels, compactness index, and regularity metrics in different methods. The curious result is that no performing-best-in-all metrics appeared, and expected high performance in one metric would necessitate a specific algorithm.

[12] Arguably one of the most used earth observation satellite constellation, from European Space program's Earth Observation component of Copernicus, providing 10 meters x 10 meters spatial resolution, close to one week revisit time for most of the world and publicly available for both non-commercial and commercial uses

Last week in Environmental Chemistry

Author: Yasin Güray Hatipoğlu

These are from chemRxiv's Earth, Space, and Environmental Chemistry categories in the September 2-6, 2024 period.

Goswami et al.[24] studied the transformation products of antimicrobials under fenton-enhanced[13] zero-valent iron powder oxidation. The organic compounds they studied were sulfamethoxazole (antibiotic), gabapentin (drug), diuron (a herbicide), terbutryn and terbuthlazine (herbicides). Zero-valent iron and Fenton oxidation process batches were mostly in room conditions in beakers. Solution pH was adjusted to 5 before iron addition, and different pollutants had varying optimum reaction time and hydrogen peroxide:iron (H_2O_2:Fe(0)) ratios. To detect analytes and transformation products, liquid chromatography-mass spectrometry (LC/MS) was utilized, and as a control of the reaction impact, Milli Q water was also treated with the same oxidation experiment, and resulting MS peaks were subtracted from the experimental results. They presented the LC-MS method parameters for these experiments as well. An electrospray Ionization detector was used at the end. For sulfamethoxazole specifically, the rate of change of transformation products was measured by taking samples at different times after the reaction initiation. They also utilized an Ecological Structure Activity Relationship with the ECOSAR program of the US EPA, which uses quantitative structure-activity relationships. Except for Gabapentin with 65 to 70 % degradation and Diuron with 80 to 90 % degradation, other chemicals were completely oxidized to their transformation products. They estimated low aquatic toxicity of transformations and reported Fenton reaction potential attack locations as the aliphatic chains of the heteroatom.

Trueman et al.[25] developed a comprehensive approach for environmental data analysis, especially focusing on how to handle non-detections. Current common approaches to non-detections by imputing them with a fixed value or completely removing them jeopardize the entire data analysis. This cutoff, generally known as the limit of detection, or for some cases, a higher value the limit of quantification, leads to the left-censored data (the censor is that rather than a statistical distribution with tails, like a normal distribution, there is a sudden, abrupt end, coming from the method's detection limit instead of the actual environmental value). Authors put forward several alternatives to impute missing data and, discussing the context of environmental samplings, like different locations, time-series, they recommend Bayesian methods. They used several metal concentrations in municipal biosolids[14] from three wastewater treatment plants. They used R version 4.3.3 for these analyses. They first showed the severe negative consequence of replacing non-detections with a constant value. The success of the Bayesian approach was evident in modeling the cumulative distribution function of the left censored data with a linear equation, and an even more complex continuous time autoregression-based smoothed spline was applied for Titanium concentrations later. They reported an improvement in the prediction as the previous time's concentration helped in better prediction. They also explored one-step multiple imputation with Bayesian correlation matrix and probabilistic principal component analysis and interpreted their results.

[13] Fenton Reaction is a highly oxidizing advanced oxidation process type, where hydrogen peroxide decomposition and hydroxyl radical generation was "catalyzed". Even though it is stated as catalyst, it transform into another species, mostly ferric cation or precipitate as oxides, and for advanced oxidation, comparatively large amounts of iron is used. Normally it is done with ferrous cations. Later, many different variants have been developed involving different cations or forms of iron with auxiliary agents.

[14] municipal wastewater sludge

Last week in Data Science-Applied Statistics

Author: Yasin Güray Hatipoğlu

These preprints are from arXiv's stat.ML and stat.AP categories in the September 2-6, 2024 period.

Wang[26] examined new ways to develop a better-performing discriminant analysis by combining Pillai's trace with uncorrelated linear discriminant analysis. Wang established this forward discriminant analysis framework as traditional linear discriminant analysis (LDA) was found to be sensitive to noise and computationally problematic in several cases. LDA first prepares within-group and between-group differences, and then class-based scatter matrices from these values are created. Later, Wilks' Λ is the division of the within-class scatter matrix to the total scatter matrix. The main purpose is to create features that maximize between-groups separation and minimize within-group separation. One problem of LDA algorithms is that there can be many instances where one side of the equation is $0/0$ and these result in premature stop. Another one is the F score and hypothesis testing assumptions, that F-statistics does not follow F-distribution and variables are not ordered randomly, but selected in a stepwise manner as opposed to what was assumed while using the technique. Furthermore, a variable number of independent fixed thresholds for F-statistics was stated to result in an inflated Type I Error Rate. ULDA, first, accounted for noninvertible within-class scatter matrix problems using between-class and total scatter matrices with the Moore-Penrose inverse method. This approach is robust against noninvertible matrices, and Pillai's trace is better than Wilks' Λ as it does not override 0 to remaining cases when a zero appears at once. Wang elaborated on the algorithm and also reported its performance according to the Type I Error. Both simulated data and Iris dataset were studied and algorithms pseudocodes were provided clearly.

Pagliardini et al.[27] considered a better alternative to Exponential Moving Average (EMA) gradients. The optimization procedure makes the previous gradients redundant in progress, hence moving average-like utilization of recent cases and discarding older ones is applied. Hence, selecting how much of the old and new gradients are to be retained and how long is another optimization process, and one EMA is insufficient in this task. Their novel optimizer AdEMAMix, mixes two EMA with Adam optimizer (Adaptive Method) to better utilize past gradients without jeopardizing new gradients. They made a specific literature review on deep learning optimizers, works using additional momentum terms, and distributed optimization, and provided their novel algorithm with a theoretical basis and pseudocode. As a result, AdEMAMix large language model (LLM) trained on 101B tokens performed comparably to an AdamW model on 197B tokens, while also slowing down model forgetting in training.

Kassi and Wang[28] focused on anisotropy and its directional nature, especially for multivariate functional data, and provided a rigorous theoretical basis. Then, to showcase numerical properties, they used the R Statistical Software package direg for simulating a class of bivariate anisotropic processes. Their two processes were fractional Brownian motions. They also provided computational complexity with *O()* notations, and to the best of their knowledge, the functional data analysis domain has yet to have any other algorithm with a better computational complexity than theirs.

Lafargue et al.[29] stressed the confidence interval calculation problem in few-shot learning and recommended a new way. The main problem is the appearance of the same samples in different tasks and cases. Confidence intervals consider the sampler randomness but the data can be the same as in maybe several other task tests. Authors separated this case as Closed Confidence Intervals since they keep using the data same as in a *closed* set, the same sample, while ML practitioners are usually interested in **different** data coming from the same underlying distribution, as defined Open Confidence Interval in this study. They conducted different tests with feature extractors CLIP and DINO and showed that OCI is generally much larger than CCI, and practically OCI is nearer to the expected performance with other data than what the model was constructed on.

Cardoso et al.[30] discussed the limited reporting capability of confusion matrix-related metrics and studied ways Item Response Theory concepts can enrich their outputs. They used the Heart-Statlog dataset with 13 features and 270 instances and with a compendium of 20 different hyperparameter settings from 10 models, they had 200 models analyzing the input data separated in training-test and providing results. These models were decision tree, random forest (RF), adaptive boosting (ADA), gradient boosting, bagging, multilayer perceptron, k-nearest neighbors, support vector machine, linear support vector machine, and linear discriminant analysis. Item Response Matrix includes the probability of a correct answer for fine-tuned models, and this was later considered with a confusion matrix. They employed Scikit-learn in Python for this process. Item parameter histogram gives information on

the discriminative power of instances, the difficulty of correct classification, and how many of instances are being classified by chance. In the end, choosing the most appropriate model for a given dataset obligates considering the complexity of the dataset and instance-level examination according to the features.

Jinng and Liu[31] emphasized the problem that arose in algorithm benchmarking from the Non-Independence from Irrelevant Alternatives - NIIA. A and B algorithms are tested and found A is superior, then a C algorithm is also tested but after this test, the previously superior A algorithm is found inferior to the B algorithm. This is not an intuitively expected result and should not be the case. In their toy dataset, they showed that non-parametric hypothesis testing and Bayesian Inference both had this NIIA problem. In such cases, they tracked down the root of NIIA and presented the information loss after rank normalization. This is defined as the *erasure effect of rank normalization*. They recommended absolute ranking prior to data analysis.

Spatial Data

Zhang et al.[32] generated a hybrid framework in spatial interpolation from scattered point sampled data via utilizing domain knowledge. Common ways to interpolate sampled points (inverse distance weighting [IDW], Kriging) require very simple constraints, such as assuming sufficient samples in all cases, linear changes, and many other limits. Their method is in two steps, the first one is data-driven Spatial Dependency Basis extraction, and the second one is rule-assisted spatial dependency approximation. Adaptive-Network-based Fuzzy Inference System (ANFIS) was stated to be intractable when there are so many rules, and this study mitigates this with *input feature decomposition*. Basically, they create a spatial dependency basis from observations, then use these so separate expert judgment-based rules and work through ANFIS. [15]. Their pseudocode is available, and they worked through a case of oil reservoir porosity mapping after random sampling and interpolating the map again. They also applied their method to ozone measurements from US Environmental Protection Agency (EPA) air quality monitoring stations. Especially in absolute error and absolute percentage error metrics, their method outperforms ordinary methods, IDW, and Kriging in both cases.

[15] The structure of the rules can allow non-linearity capture, as in decision trees, piecewise linear regression, or the like, albeit with specific rule sets defined *a priori*.

Time-series Data

Xie et al.[33] stated the overall success in finding how the data changes by common time-series analysis approaches like auto-regressive integrated moving average (ARIMA), Kalman filter, gradient boosting decision tree, recurrent neural networks, temporal convolutional networks, and X-formers, they also stressed their limitation in not giving the factors affecting this dataset. Instead, they put symbolic regression forward but stated their computational inefficiency and complex heuristics. As a result, they utilize Neural-Enhanced Monte-Carlo Tree Search (NeMCTS) that overcomes the limitations of symbolic regression. MCTS backbone has the selection, expansion, simulation, and back-propagation steps. The goal is to generate an analytical function for the given time series, and its reward structure considers both the accuracy of the value in every timestamp and the simplicity of the expression in general (the simpler better). In model training, the policy value network utilized a Long Short-Term Memory network for expression path encoding and temporal convolutional networks for input signal processing. They used weighted influenza-like illness percentage (WILI), Australian Daily Currency Exchange Rates (ACER), and atmospheric pressure (AP) datasets. The performance comparison of this novel method was done with Genetic Programming, Multiple Regression Genetic Programming, Batesian Symbolic Regression, Physics Symbolic Optimization, and Symbolic Physics Learner. NeMCTS outperformed other algorithms in ACER and AP datasets in all metrics. In WILI, it was quite near to the Symbolic Physics Learner, the best performer, in terms of correlation coefficient, coefficient of determination R^2, and average time cost per sample for time to finish the run. Extrapolation (all best but AP correlation coefficient) and Ablation studies were conducted.

References

[1] Tom Bakx and John Conway. Recommended receiver papers for alma users, 2024, 2409.02164.

[2] Jiaming Wang, William H. Matthaeus, Rohit Chhiber, Sohom Roy, Rayta A. Pradata, Francesco Pecora, and Yan Yang. $1/f$ noise in the heliosphere: A target for punch science, 2024, 2409.02255.

[3] L. Colzi, V. M. Rivilla, M. T. Beltrán, C. Y. Law, E. Redaelli, and M. Padovani. Astrochemistry on galactic scales, 2024, 2409.02537.

[4] Thai Duy Quy, Alvin Buana, Josh Lee, and Rakha Asyrofi. Hazardous asteroids classification, 2024, 2409.02150.

[5] Keming Zhang, Weicheng Zang, Kareem El-Badry, Jessica R. Lu, Joshua S. Bloom, Eric Agol, B. Scott Gaudi, Quinn Konopacky, Natalie LeBaron, Shude Mao, and Sean Terry. An earth-mass planet and a brown dwarf in orbit around a white dwarf, 2024, 2409.02157.

[6] Jessica Speedie, Ruobing Dong, Cassandra Hall, Cristiano Longarini, Benedetta Veronesi, Teresa Paneque-Carreño, Giuseppe Lodato, Ya-Wen Tang, Richard Teague, and Jun Hashimoto. Gravitational instability in a planet-forming disk. *Nature*, 633(8028):58–62, September 2024.

[7] Elise L. Evans, Trent J. Dupuy, Kendall Sullivan, Adam L. Kraus, Daniel Huber, Michael J. Ireland, Megan Ansdell, Rajika L. Kuruwita, Raquel A. Martinez, and Mackenna L. Wood. Orbital architectures of planet-hosting binaries iii. testing mutual inclinations of stellar and planetary orbits in triple-star systems, 2024, 2409.02223.

[8] David R. Coria, Neda Hejazi, Ian J. M. Crossfield, and Maleah Rhem. The Wanderer: Charting WASP-77A b's Formation and Migration Using a System-Wide Inventory of Carbon and Oxygen Abundances. *arXiv e-prints*, page arXiv:2409.02286, September 2024, 2409.02286.

[9] D. Föhring, L. Conversi, M. Micheli, E. Dölling, and P. Ramirez Moreta. Site selection for the second flyeye telescope: A simulation study for optimizing near-earth object discovery. *Icarus*, 424:116281, December 2024.

[10] Tanvi Goswamy, Andrew Collier Cameron, and Thomas G. Wilson. Do anomalously-dense hot jupiters orbit stealth binary stars?, 2024, 2409.02639.

[11] F. Castelli and R. L. Kurucz. New Grids of ATLAS9 Model Atmospheres. In N. Piskunov, W. W. Weiss, and D. F. Gray, editors, *Modelling of Stellar Atmospheres*, volume 210 of *IAU Symposium*, page A20, January 2003, astro-ph/0405087.

[12] Joshua J. Brown and Gordon I. Ogilvie. Horseshoes and spiral waves: capturing the 3d flow induced by a low-mass planet analytically, 2024, 2409.02687.

[13] Julien Laurent-Varin, Théo James, Jean-Charles Marty, Laurent Jorda, Sebastien Le Maistre, and Robert Gaskell. New gravity field of comet 67p/c-g based on rosetta's doppler and optical data, 2024, 2409.02692.

[14] Sota Arakawa, Daiki Yamamoto, Lily Ishizaki, Tamami Okamoto, and Noriyuki Kawasaki. Oxygen isotope exchange between dust aggregates and ambient nebular gas, 2024, 2409.02736.

[15] Thomas Meier, Christian Reinhardt, Miles Timpe, Joachim Stadel, and Ben Moore. A systematic survey of moon-forming giant impacts. ii. rotating bodies, 2024, 2409.02746.

[16] Christophe Lovis, Nicolas Blind, Bruno Chazelas, Muskan Shinde, Maddalena Bugatti, Nathanaël Restori, Isaac Dinis, Ludovic Genolet, Ian Hughes, Michaël Sordet, Robin Schnell, Samuel Rihs, Adrien Crausaz, Martin Turbet, Nicolas Billot, Thierry Fusco, Benoît Neichel, Jean-François Sauvage, Pablo Santos Diaz, Mathilde Houelle, Joshua Blackman, Audrey Lanotte, Jonas G. Kühn, Janis Hagelberg, Olivier Guyon, Patrice Martinez, Alain Spang, Christoph Mordasini, David Ehrenreich, Brice-Olivier Demory, and Emeline Bolmont. Ristretto: reflected-light exoplanet spectroscopy at the diffraction limit of the vlt. In Joël R. Vernet, Julia J. Bryant, and Kentaro Motohara, editors, *Ground-based and Airborne Instrumentation for Astronomy X*, page 54. SPIE, July 2024.

[17] MD Ridwan Adib, Colm Barrett, Shane O'Sullivan, Anna Flynn, Marie McFadden, Emer Kennedy, and Alan O'Riordan. In situ ph-controlled electrochemical sensors for glucose and ph detection in saliva, 2024.

[18] Allison B. Esselman, Megan S. Ward, Cody R. Marshall, Martin Dufresne Ellie L. Pingry, Melissa A. Farrow, Matthew Schrag, and Jeffrey M. Spraggins. A streamlined workflow for microscopy-driven maldi imaging mass spectrometry data collection, 2024.

[19] Nathan Heath Patterson, Elizabeth K. Neumann, Kavya Sharman, Jamie Allen, Raymond Harris, Agnes B. Fogo, Mark de Caestecker, Richard M. Caprioli, Raf Van de Plas, and Jeffrey M. Spraggins. Autofluorescence microscopy as a label-free tool for renal histology and glomerular segmentation. *bioRxiv*, 2021.

[20] Kentarou Fuku and Takefumi Yoshida. Unsupervised machine learning-based image recognition of raw ir spectra: Toward chemist-like chemical structural classification, 2024.

[21] Thanh-Dung Le, Vu Nguyen Ha, Ti Ti Nguyen, Geoffrey Eappen, Prabhu Thiruvasagam, Luis M. Garces-Socarras, Hong fu Chou, Jorge L. Gonzalez-Rios, Juan Carlos Merlano-Duncan, and Symeon Chatzinotas. On-board satellite image classification for earth observation: A comparative study of pre-trained vision transformer models, 2024, 2409.03901.

[22] Isaac Ray and Alexei Skurikhin. Deep clustering of remote sensing scenes through heterogeneous transfer learning, 2024, 2409.03938.

[23] Isabela Borlido, Eduardo Bouhid, Victor Sundermann, Hugo Resende, Alvaro Luiz Fazenda, Fabio Faria, and Silvio Jamil F. Guimarães. How to identify good superpixels for deforestation detection on tropical rainforests, 2024, 2409.04330.

[24] Anuradha Goswami, Jia-Qian Jiang, Roberts Joanne, and Michael Petri. Fenton-enhanced zero-valent iron powder oxidation: Investigating transformation products of antimicrobials and its removal, 2024.

[25] Benjamin F. Trueman, Madison Gouthro, Amina K. Stoddart, and Graham A. Cagnon. A comprehensive approach to analyzing environmental data with non-detects, 2024.

[26] Siyu Wang. A new forward discriminant analysis framework based on pillai's trace and ulda, 2024, 2409.03136.

[27] Matteo Pagliardini, Pierre Ablin, and David Grangier. The ademamix optimizer: Better, faster, older, 2024, 2409.03137.

[28] Omar Kassi and Sunny G. W. Wang. Structural adaptation via directional regularity: rate accelerated estimation in multivariate functional data, 2024, 2409.00817.

[29] Raphael Lafargue, Luke Smith, Franck Vermet, Mathias Löwe, Ian Reid, Vincent Gripon, and Jack Valmadre. Oops, i sampled it again: Reinterpreting confidence intervals in few-shot learning, 2024, 2409.02850.

[30] Lucas Felipe Ferraro Cardoso, José de Sousa Ribeiro Filho, Vitor Cirilo Araujo Santos, Regiane Silva Kawasaki Frances, and Ronnie Cley de Oliveira Alves. Standing on the shoulders of giants, 2024, 2409.03151.

[31] Yunpeng Jinng and Qunfeng Liu. Absolute ranking: An essential normalization for benchmarking optimization algorithms, 2024, 2409.04479.

[32] Cong Zhang, Shuyi Du, Hongqing Song, and Yuhe Wang. A hybrid framework for spatial interpolation: Merging data-driven with domain knowledge, 2024, 2409.00125.

[33] Yi Xie, Tianyu Qiu, Yun Xiong, Xiuqi Huang, Xiaofeng Gao, and Chao Chen. An efficient and generalizable symbolic regression method for time series analysis, 2024, 2409.03986.