

УДК 532.582.32, 004.942, 004.67

МОДЕЛИРОВАНИЕ ХАРАКТЕРИСТИК ТОНАЛЬНОГО ШУМА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ ПРИ ОБТЕКАНИИ ГИДРОПРОФИЛЯ

© С. С. Абдуракипов^{1,2}, М. П. Токарев^{1,2},
К. С. Первунин^{1,2}, В. М. Дулин^{1,2}

¹Институт теплофизики СО РАН,
630090, г. Новосибирск, просп. Академика Лаврентьева, 1
²Новосибирский государственный университет,
630090, г. Новосибирск, ул. Пирогова, 2
E-mail: s.s.abdurakipov@gmail.com

Апробирован метод машинного обучения для моделирования характеристик тонального шума, возникающего при обтекании гидропрофиля. На основе данных эксперимента проведены исследования математических моделей регрессии амплитуды давления и моделей классификации режимов с высоким уровнем тонального шума от безразмерных параметров обтекания. Протестированы различные семейства алгоритмов: от линейных до нейросетевых. Показано, что реализованная модель градиентного бустинга с коэффициентом детерминации (95 %) является наиболее точной для описания спектральных кривых акустического давления на всём интервале значений амплитуд и характерных частот.

Ключевые слова: машинное обучение, обтекание гидропрофиля, тональный шум.

DOI: 10.15372/AUT20190213

Введение. Известно, что формирование акустического шума приводит к понижению эффективности летательных аппаратов, а также увеличивает шумовую загрязнённость вблизи крупных аэропортов, что негативно сказывается на здоровье людей, поэтому исследование данного процесса важно для аэрокосмической промышленности. Основной составляющей общего шума летательного аппарата является собственный шум аэродинамического профиля, который обусловлен взаимодействием профиля с турбулентными пульсациями от вихревых структур, формирующимися в пограничном слое и в среде обтекания. Предложен ряд подходов к математическому моделированию характеристик собственного шума профиля. Согласно [1] все подходы можно разделить на три группы: аналитическая теория, основанная на акустической аналогии Лайтхилла [2], аппроксимация линеаризованными уравнениями гидромеханики [3] и полуэмпирические модели [4]. В [4] определены пять режимов обтекания с генерацией тонального шума из-за вихревых явлений, характерных для пограничного слоя: 1 — с турбулентными пульсациями пограничного слоя, проходящего через задний край гидропрофиля, 2 — с турбулентными пульсациями в потоке с отрывом пограничного слоя и при формировании глубокого свала, 3 — с формированием вихрей вследствие неустойчивости ламинарного пограничного слоя, 4 — с формированием вихрей от тупых задних кромок и 5 — с турбулентными пульсациями от вихрей вблизи законцовок лопастей несущего винта. Общий уровень акустического давления складывается из всех описанных выше механизмов, которые зависят от условий обтекания аэродинамического профиля. Модель Брукса, Поупа и Марколини [4] построена на основе большой базы данных испытаний акустической аэродинамической трубы, выполненных NASA на различных профилях аэродинамической поверхности NASA 0012. Полуэмпирическая модель Брукса широко применяется в настоящее время для моделирования амплитуды

тонального шума аэродинамического профиля при натекании ламинарного потока в зависимости от угла атаки, скорости набегающего потока и геометрических параметров аэродинамического профиля. Однако модель Брукса не позволяет точно предсказать уровень акустического давления, особенно наблюдается недооценка низких уровней акустического давления. Это указывает на необходимость разработки более точной математической модели для описания тонального шума, создаваемого аэродинамическим профилем. Развитие технологий машинного обучения, основанных на градиентном бустинге, нейросетевых алгоритмах, имеет перспективу для создания улучшенной математической модели характеристик тонального шума, которую можно использовать на практике. Существует большое количество примеров использования методов машинного обучения в ряде авиационных задач моделирования эмпирических закономерностей, диагностики неисправностей и управления [5].

Цель данной работы — проанализировать применимость и реализацию моделей амплитудно-частотных кривых тонального шума при обтекании гидропрофиля на основе развитых алгоритмов машинного обучения. Задача рассматривается в двух постановках: исследование модели регрессии амплитуды давления и модели классификации режимов с высоким уровнем тонального шума.

Экспериментальные измерения. Панорамные измерения полей мгновенной скорости методом PIV (Particle Image Velocimetry) и измерения пульсаций давления (датчиком давления) выполнялись на кавитационном стенде в Институте теплофизики СО РАН. Подробное описание стенда и методики измерений приведено в работе [6]. Экспериментальные исследования проводились в целях изучения характеристик турбулентных пульсаций скорости и давления при обтекании гидропрофиля в условиях формирования вихревых структур. Анализ динамики и характеристик вихревых структур и их влияние на формирование акустического шума будут рассмотрены далее.

Для построения модели использовалась открытая база данных NASA [4], дополненная собственными экспериментальными измерениями. Набор данных NASA включает аэродинамические профили NASA 0012 различного размера при разных скоростях обтекания в аэродинамической трубе и углах атаки. В проведённых лабораторных измерениях применялся профиль NASA 0015, который имеет близкие к NASA 0012 характеристики обтекания. В процессах измерений были получены спектры тонального шума профилей. Все результаты представлены в виде 1/3 октавных спектров. Таким образом, общий набор данных состоит из 1755 записей и следующих переменных: относительного давления P (дБ), частоты f (Гц), угла атаки α ($^\circ$), длины хорды l (м), скорости набегающего потока U (м/с), толщины пограничного слоя на стороне разряжения Δ (м). Все физические величины были преобразованы в безразмерные гидродинамические параметры: число Рейнольдса $Re = Ul/\nu$ (ν — коэффициент кинематической вязкости), толщина пограничного слоя $\delta = \Delta/l$, число Струхала $St = f\delta/U$, $St_l = fl/U$, число Маха $Ma = U/c$ (c — скорость звука), число Рошко $Ro = StRe$.

Описание методов моделирования. В данной работе проанализированы различные семейства моделей: от линейных алгоритмов до нейросетевых. Задачей машинного обучения является восстановление неизвестной зависимости $y(x)$ по эмпирическим данным с учётом признакового описания (параметров) конечного набора наблюдений. Линейные модели реализуют решение на основе линейной комбинации переменных (признаков) x_i , взятых с весами w_i : $a(w, x) = \sum_i w_i x_i$. Поиск оптимальных весов обычно сводят к задаче минимизации функции ошибки методом наименьших квадратов: $Q(X, w) = \min_w \|Xw - y\|_2^2$ [7], где X — матрица наблюдений, y — вектор целевой переменной, w — вектор весов модели. Гребневая регрессия и Лассо [7] — это модификации метода наименьших квадратов, заключающиеся в использовании регуляризации, т. е. наложении штрафа

за большую норму (l_1 -регуляризация) или квадрат нормы весов (l_2 -регуляризация). Регуляризация применяется для упрощения сложности модели и предотвращения переобучения, когда ошибка модели на тестовой выборке оказывается существенно выше средней ошибки на обучающей выборке [7]. В уравнении $w = \operatorname{argmin} \|Xw - y\|_2^2 + \alpha\|w\|_2^2 + \beta\|w\|_1$ коэффициенты α и β являются параметрами регуляризации. Метод опорных векторов для задачи регрессии заключается в переводе исходных векторов в пространство более высокой размерности и поиске аппроксимирующей поверхности с максимальным зазором в этом пространстве. Преимуществом метода является возможность его использования в спрямляющем пространстве посредством ядерных функций, что позволяет описывать нелинейные закономерности [7]. Математически уравнение выглядит как

$$a(w, x) = \sum_i ((\alpha_i - \alpha_i^*)K(x, x_i) + b),$$

где α_i — веса переменных, $K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / (2\sigma^2)}$, $K(x_i, x_j) = (x_i, x_j)^d$ — радиальная и полиномиальная ядерные функции.

Ансамбль деревьев (случайный лес деревьев) — это множество независимых базовых моделей «деревьев решений». Для построения независимых деревьев используется метод случайных подпространств, который позволяет снизить коррелированность между деревьями и избежать переобучения [8]. В методе случайных подпространств базовые алгоритмы (деревья) обучаются на различных подмножествах признакового описания, которые также выделяются случайным образом. В задачах регрессии обычно применяют «жадный» подход к построению деревьев (последовательно от «корня» к «листьям») и минимизируют дисперсию целевой переменной в «листьях» деревьев. Результатом ансамбля деревьев является усреднение всех базовых моделей.

Градиентный бустинг представляет алгоритм как направленную композицию более простых моделей ($a(x) = \sum_i \beta_i h_i(x)$) и при этом использует решающие деревья фиксированного размера в качестве базовых алгоритмов $h_i(x)$. «Деревья решений» имеют ряд преимуществ, например способность обрабатывать данные смешанного типа и моделировать сложные нелинейные функции. На каждой итерации градиентного бустинга модель улучшается «жадным» образом: $a_i(x) = a_{i-1}(x) + \beta_i h_i(x)$, причём каждый новый алгоритм уменьшает значение функции потерь $L(y_i, a_{i-1}(x_i))$ ансамбля на предыдущей итерации. Коэффициент β называется шагом градиентного бустинга (скоростью обучения). Задача решается методом градиентного спуска:

$$h_i(x) = \operatorname{argmin}_{h(x)} \sum_i (L(y_i, a_{i-1}(x_i)) + h(x_i)).$$

Детали используемых методов можно найти в [8].

Далее будут использованы сокращения: Ridge — гребневая регрессия, KNN — метод ближайших соседей, LR — логистическая регрессия, SVM L — линейный метод опорных векторов, SVM NL — метод опорных векторов с нелинейным ядром, DT — «дерево решений», LAR — Лассо-регрессия наименьшего угла, BR — байесовская гребневая регрессия, SVR L — линейный метод опорных векторов (регрессия), SVR NL — метод опорных векторов с нелинейным ядром (регрессия), RF — случайный лес деревьев решений, GB — градиентный бустинг над «деревьями решений», MLP — многослойный перцептрон (нейронная сеть).

Результаты моделирования. В данной работе проведено исследование возможности построения математической модели регрессии амплитуды давления и модели классификации режимов с высоким уровнем тонального шума. На рис. 1 для примера показаны анализируемые амплитудно-частотные (P - St) кривые относительного акустического

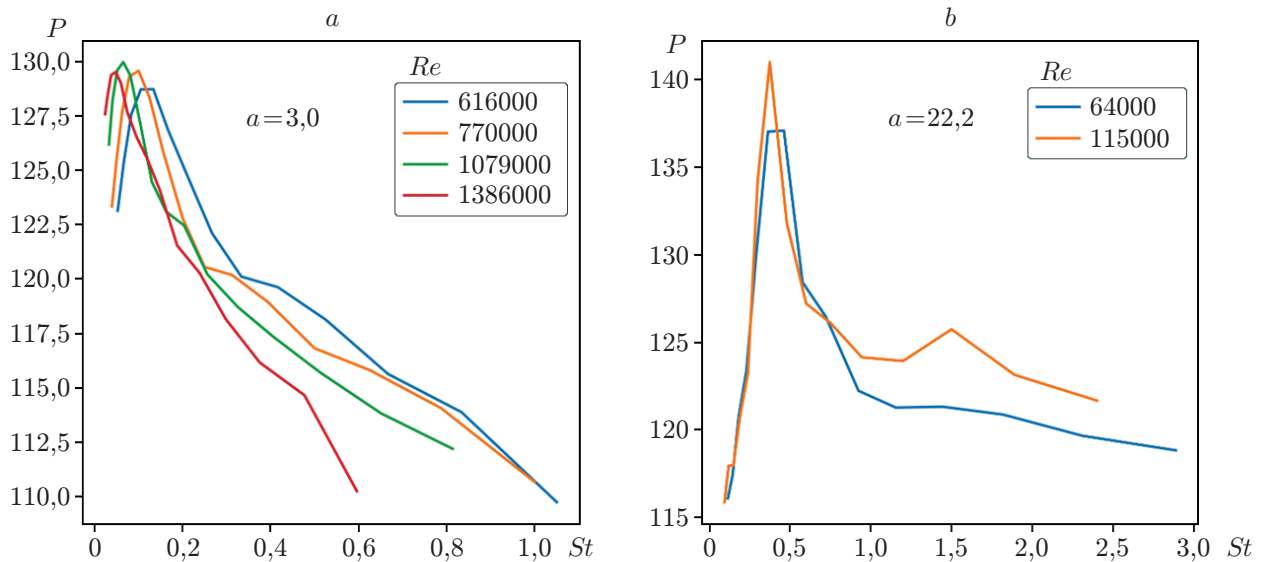


Рис. 1. Зависимость относительной амплитуды давления P (дБ) от безразмерной частоты (St) для различных чисел Рейнольдса и двух углов атаки: $\alpha = 3,0$ (а) и $\alpha = 22,2$ (б)

Таблица 1

Результаты тестирования моделей классификации на кроссвалидации, где для каждой модели $a_n(x)$ показаны среднее значение μ , стандартное отклонение σ , минимальное m и максимальное M значения метрики качества AUC ROC на 10 блоках

$a_n(x)$	μ	σ	m	M
Ridge	0,8715	0,024	0,8216	0,9079
KNN	0,9512	0,02	0,9113	0,9801
LR	0,8855	0,021	0,8657	0,9374
SVM L	0,8796	0,022	0,8388	0,9283
SVM NL	0,9720	0,009	0,9595	0,9888
DT	0,9388	0,019	0,9123	0,9643
RF	0,9773	0,013	0,9437	0,9920
GB	0,9790	0,015	0,9426	0,9957
MLP	0,9529	0,0138	0,923	0,9692

давления P (дБ) при различных числах Рейнольдса (Re) для малого $\alpha = 3,0$ и большого $\alpha = 22,2$ углов атаки. В процессе исследования протестированы различные семейства моделей $a_n(x)$: от линейных до нейросетевых алгоритмов. Алгоритмы реализованы на языке Python с применением библиотек Scikit-learn, Keras, Hyperopt, Shar. Статистическая выборка, полученная из данных эксперимента, была разбита на обучающую (83 %) и тестовую (17 %) подвыборки. Обучающая часть затем была поделена на 10 блоков с перемешиванием наблюдений для осуществления процедуры кроссвалидации [8], с помощью которой подбирались оптимальные гиперпараметры моделей. Качество лучшей модели затем оценивалось на тестовой выборке. В табл. 1 приведены результаты оценки качества моделей классификации наблюдений с высоким уровнем шума (выше среднего уровня). Качество моделей оценивалось по метрике AUC ROC (площадь под кривой ROC, Area Under Curve of Receiver Operating Characteristic), которая не зависит от выбора оптимальной отсечки по вероятности принадлежности к классу [8]. Можно видеть, что метод опорных векторов с нелинейным ядром ($\mu_{AUC} = 0,972$), случайный лес ($\mu_{AUC} = 0,9773$) и гради-

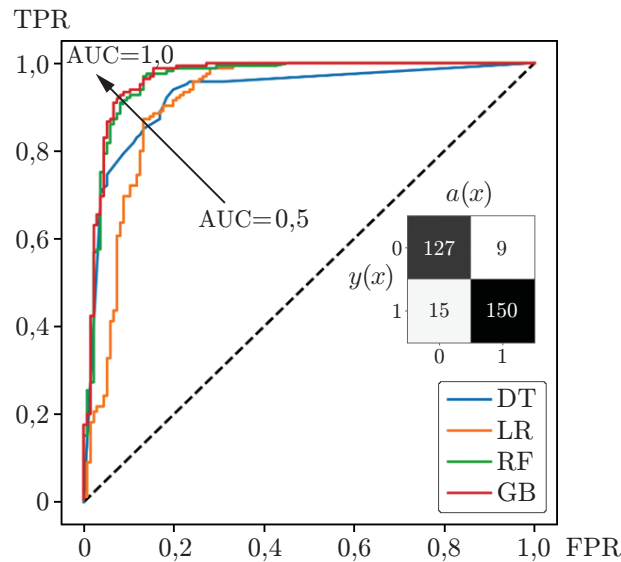


Рис. 2. ROC-кривые для ряда моделей и матрица ошибок для лучшей модели градиентного бустинга; в матрице ошибок $y(x)$ — факт, $a(x)$ — результат модели

ентный бустинг ($\mu_{\text{AUC}} = 0,979$) демонстрируют наилучшие результаты классификации при близких средних показателях метрики. На рис. 2 визуализированы ROC-кривые (зависимость чувствительности, true positive rate (TPR), модели от 1-специфичности, false positive rate (FPR)) и матрица ошибок лучшей модели градиентного бустинга (GB) [8]. Модель тестировалась на 301 наблюдении (17 % выборки) и показала 15 и 9 ошибок первого и второго рода соответственно. Таким образом, точность определения наблюдений с высокими показателями давления составила 92 %.

На рис. 3, a представлена относительная частота n_t/N , с которой каждая переменная встречается в ансамбле деревьев градиентного бустинга. Данный показатель качественно свидетельствует о степени важности переменной для модели [8]. Чтобы более глубоко проанализировать переменные, которые являются наиболее важными для построенной математической модели, необходимо рассчитать значения Sh (shapley additive explanations) переменных для каждого объекта в тестовой выборке [9]. Величина Sh , характеризующая степень важности переменной для конкретного наблюдения в статистической выборке, представляет собой средний вклад в целевую функцию от значения признака, добавленного к совокупности других признаков (при всех возможных перестановках) [9]. На рис. 3, b показана зависимость вклада Sh признака x в значение целевой функции каждого наблюдения в тестовой выборке [9]. Цвет отражает величину переменной (синий — минимальное значение, красный — максимальное), а вертикальная координата пропорциональна относительной частоте, с которой в выборке встречается данное значение. Так, например, высокие значения числа Рощко (Ro) и числа Струхала (St , St_l), а также толщины пограничного слоя (δ) и угла атаки (α) снижают амплитуду тонального шума для большого числа наблюдений.

При решении задачи регрессии особое внимание было уделено выявлению дополнительных переменных, которые могли бы повысить качество моделей и иметь физический смысл. Для формирования новых признаков использовались полиномиальные параметры и их попарные произведения, так как степенные зависимости безразмерных гидродинамических параметров зачастую входят в основные уравнения гидромеханики. Отбор признаков осуществлялся посредством процедуры «жадного» отбора на основе модели (ансамбля

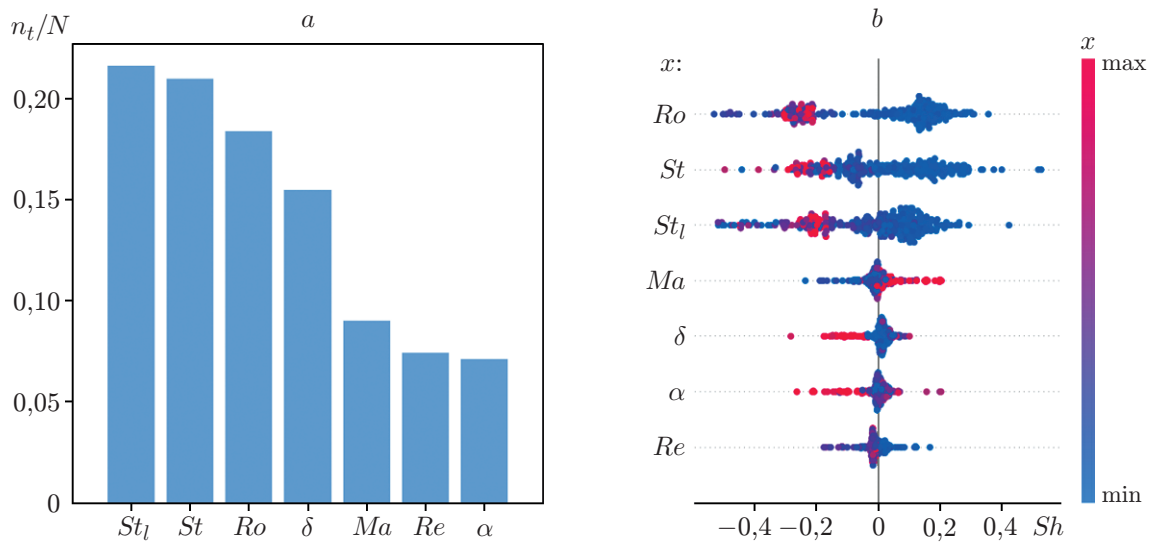


Рис. 3. Анализ зависимостей уровня шума от гидродинамических параметров: a — относительная частота, с которой переменная встречается в ансамбле деревьев градиентного бустинга; b — зависимость вклада Sh признака x в значение целевой функции каждого примера в тестовой выборке [9] (цветом показана величина переменной)

Таблица 2

Результаты тестирования моделей регрессии на кроссвалидации, где для каждой модели $a_n(x)$ показаны среднее значение μ , стандартное отклонение σ , минимальное m и максимальное M значения метрики качества R^2 на 10 блоках

$a_n(x)$	μ	σ	m	M
Ridge	0,6253	0,056	0,5241	0,7179
KNN	0,8740	0,026	0,8292	0,9186
LAR	0,6220	0,051	0,5247	0,7102
BR	0,6252	0,055	0,5247	0,7174
SVR L	0,6102	0,061	0,4995	0,6809
SVR NL	0,8238	0,045	0,7597	0,8826
DT	0,8156	0,035	0,7609	0,8687
RF	0,9000	0,021	0,8580	0,9259
GB	0,9382	0,015	0,9060	0,9587
MLP	0,8901	0,023	0,8550	0,9213

деревьев) [8]. Результаты математических моделей регрессии на отобранном наборе наиболее влияющих переменных можно видеть в табл. 2. Для оценивания качества и подбора гиперпараметров моделей была использована процедура кроссвалидации на 10 блоках.

Метрикой качества был выбран коэффициент детерминации R^2 . Для подбора гиперпараметров использовалась процедура байесовской оптимизации (на основе метамодели) с алгоритмом Tree Parzen Estimator (ТРЕ) из библиотеки Нургопт [10]. Наилучшим образом себя показали три алгоритма: однослойный перцептрон (нейронная сеть) ($\mu_{R^2} = 0,89$), случайный лес деревьев ($\mu_{R^2} = 0,9$) и градиентный бустинг ($\mu_{R^2} = 0,94$). Линейные модели не смогли превысить порог $\mu_{R^2} > 0,63$. Сравнение фактических значений уровня безразмерного давления P (дБ) и результатов моделирования P_m (дБ) для трёх моделей показано на рис. 4. Наименьшая дисперсия точек относительно точной прямой наблюдается

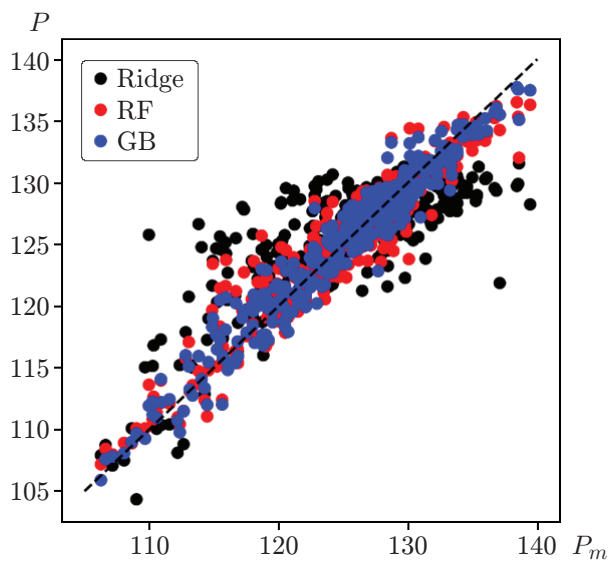


Рис. 4. Зависимости фактической (P) от предсказанной (P_m) относительных амплитуд давления на основе ряда моделей

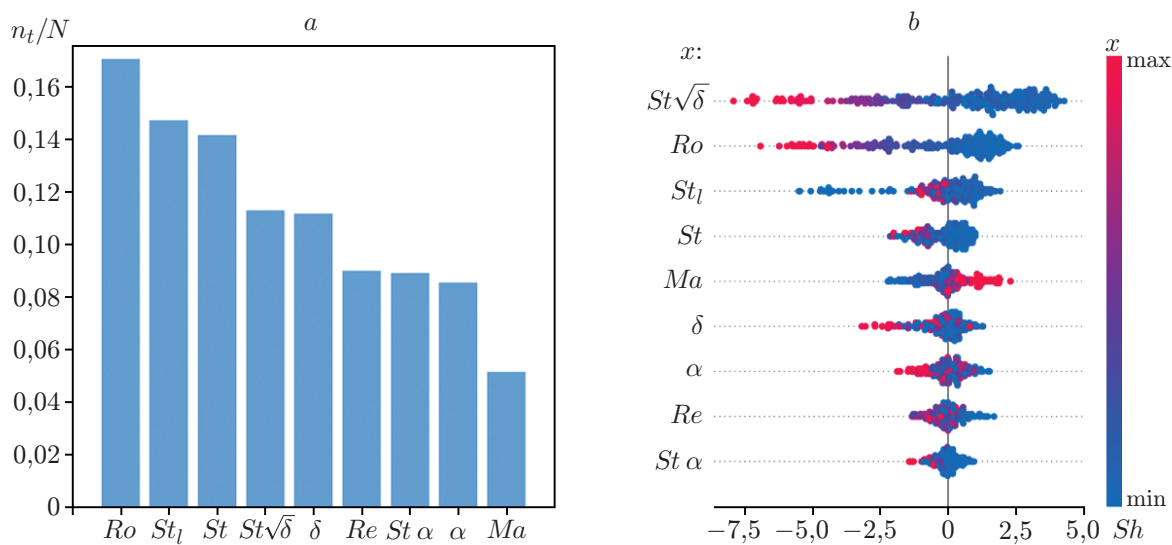


Рис. 5. Анализ зависимостей уровня шума от дополнительных гидродинамических параметров (обозначения как на рис. 3)

ся для градиентного бустинга ($R^2 = 95\%$). Использование дополнительных переменных $St\sqrt{\delta}$, $St\alpha$ и других позволило увеличить качество регрессии давления на 2–3 %. Помимо упомянутых чисел Рошко и Струхаля, толщины пограничного слоя и угла атаки комплекс $St\sqrt{\delta}$ является наиболее значимым для модели регрессии амплитуды давления большого числа статистических наблюдений (рис. 5).

Заключение. Выполненная работа направлена на исследование возможности построения математической модели прогнозирования характеристик тонального шума, возникающего при обтекании гидропрофиля, с использованием методов машинного обучения. Решены задачи определения режимов с высоким уровнем шума (классификация) и прогнозирования амплитудно-частотных кривых нормированного акустического давления в зависимости от безразмерных параметров обтекания. В целях расширения обучающей выборки данные из открытых источников были дополнены собственными экспериментальными измерениями. В процессе исследования проанализированы различные семейства статистических моделей: от линейных алгоритмов до нейросетевых. Показано, что модель градиентного бустинга над деревьями решений обеспечивает наилучшее качество при решении указанных задач. Точность решения задачи классификации составила 92 % ($AUC = 0,97$). Модель регрессии с коэффициентом детерминации ($R^2 = 95\%$) является наиболее точной для описания спектральных кривых акустического давления во всём интервале значений амплитуд и характерных частот. Проанализировано влияние различных переменных на качество моделей. Обнаружено, что толщина пограничного слоя, числа Струхаля и Рошко являются самыми важными безразмерными переменными. Результаты моделирования могут быть полезны для инженерного проектирования малошумных аэродинамических профилей. Результаты PIV-измерений поля скорости с разрешением во времени можно использовать для анализа влияния свойств когерентных вихревых структур на характеристики тонального шума с применением класса методов машинного обучения — методов понижения размерности стохастических динамических систем [11].

Финансирование. Работа выполнена при финансовой поддержке Российского научного фонда (грант № 18-79-00138).

СПИСОК ЛИТЕРАТУРЫ

1. **Howe M. S.** A review of the theory of trailing edge noise // Journ. Sound Vib. 1978. **61**, N 3. P. 437–465.
2. **Lighthill M. J.** On sound generated aerodynamically I. General theory // Proc. Roy. Soc. London. A. 1952. **211**, N 1107. P. 564–587.
3. **Amiet R. K.** Effect of the incident surface pressure field on noise due to a turbulent flow past a trailing edge // Journ. Sound Vib. 1978. **57**, N 2. P. 305–306.
4. **Brooks T. F., Pope D. S., Marcolini M. A.** Airfoil Self-Noise and Prediction. NASA Technical Report, NASA-RP-1218. NASA Reference Publication, 1989. 145 p.
5. **Failer W. E., Schrec S. J.** Neural networks: Applications and opportunities in aeronautics // Progr. Aerospace Sci. 1996. **32**, N 5. P. 433–456.
6. **Kravtsova A. Yu., Markovich D. M., Pervunin K. S. et al.** High-speed visualization and PIV measurements of cavitating flows around a semi-circular leading-edge flat plate and NACA0015 hydrofoil // Intern. Journ. Multiphase Flow. 2014. **60**. P. 119–134.
7. **Vapnik V. N.** Statistical Learning Theory. N. Y.: John Wiley, 1998. 768 p.
8. **Mohri M., Rostamizadeh A., Talwalkar A.** Foundations of Machine Learning. Cambridge: MIT press, 2012. 414 p.

9. **Lundberg S. M., Lee S. I.** A unified approach to interpreting model predictions // Proc. of the Conf. on Advances in Neural Information Processing Systems. Long Beach, USA, 4–9 Dec., 2017. P. 4768–4774.
10. **Bergstra J., Yamins D., Cox D.** Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms // Proc. of the 12th Python in Science Conf. (SCIPY 2013). Austin, Texas, 24–29 June, 2013. P. 13–20.
11. **Mullyadzhanov R. I., Sandberg R. D., Abdurakipov S. S. et al.** Propagating helical waves as a building block of round turbulent jets // Phys. Rev. Fluids. 2018. **3**, N 6. 062601.

Поступила в редакцию 30.01.2019

После доработки 14.03.2019

Принята к публикации 18.03.2019
